音频副本检测方法及装置、电子设备、计算机存储介质以及计算机程序产品

相关申请的交叉引用

本申请基于申请号为2024111429279、申请日为2024年08月20日的中国专利申请提出，并要求该中国专利申请的优先权，该中国专利申请的全部内容在此引入本申请作为参考。

技术领域

本申请涉及人工智能技术领域，尤其涉及一种音频副本检测方法及装置、电子设备、计算机存储介质以及计算机程序产品。

背景技术

音频副本检测技术是一种用于识别音频内容是否与已知音频相似或者相同的技术，若识别为是，则可将已知音频作为待识别音频的副本。音频副本检测技术通常应用于音频版权保护、音乐推荐等应用场景。相关技术在检测音频副本时，往往依赖全局音频特征的匹配，检测维度单一，容易忽略音频的细节差异，导致检测的准确性低下。如何提升音频副本检测的准确性，仍是本领域的技术人员需要研究和解决的技术问题。

发明内容

为解决上述技术问题，本申请实施例提供了一种音频副本检测方法、音频副本检测装置、电子设备、计算机可读存储介质和计算机程序产品。本申请实施例能够提升音频副本检测的准确性。

本申请实施例的一个方面，提供了一种音频副本检测方法，所述方法由电子设备执行，该方法包括：提取待处理音频对应的全局音频特征，并提取所述待处理音频对应的音频段特征序列；从音频检索库中检索所述全局音频特征对应的近似全局音频特征；获取所述近似全局音频特征所属的目标音频对应的音频段特征序列，并根据所述待处理音频和所述目标音频分别对应的音频段特征序列，确定所述待处理音频与所述目标音频之间的音频相似度；若所述音频相似度达到预设相似度以上，则将所述目标音频确定为所述待处理音频对应的副本音频。

本申请实施例的另一方面，提供了一种音频副本检测装置，该装置包括：特征提取模块，配置为提取待处理音频对应的全局音频特征，并提取所述待处理音频对应的音频段特征序列；特征检索模块，配置为从音频检索库中检索所述全局音频特征对应的近似全局音频特征，并获取所述近似全局音频特征所属的目标音频对应的音频段特征序列；相似度确定模块，配置为根据所述待处理音频对应的音频段特征序列和所述目标音频对应的音频段特征序列，确定所述待处理音频与所述目标音频之间的音频相似度；副本判定模块，配置为若所述音频相似度达到预设相似度以上，则将所述目标音频确定为所述待处理音频对应的副本音频。

本申请实施例的另一方面提供了一种电子设备，包括：一个或多个处理器；存储器，

用于存储一个或多个计算机程序，当所述一个或多个计算机程序被所述一个或多个处理器执行时，使得所述电子设备实现如前所述的音频副本检测方法。

本申请实施例的另一方面提供了一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序被电子设备的处理器执行时，使所述电子设备执行如上所述的音频副本检测方法。

本申请实施例的另一方面提供了一种计算机程序产品，包括计算机程序，该计算机程序被电子设备的处理器执行时实现如上所述的音频副本检测方法。

本申请的实施例所提供的技术方案，首先根据待处理音频的全局音频特征从音频检索库中检索出近似全局音频特征，然后根据待处理音频和目标音频各自对应的音频段特征序列，来确定出待处理音频与目标音频之间的音频相似度，可以理解为是先通过全局音频特征来反映音频的整体属性，实现较粗粒度的检测，快速缩小在音频检索库中的检测范围，提高检测效率，然后再通过音频段特征序列来捕捉音频的局部细节及时序关联，实现更细粒度的检测，使检测聚焦于潜在相似的音频，增强检测的针对性。将全局音频特征的检测和音频段特征序列的检测进行结合，减少单一特征带来的误判，由此能够保证音频副本检测的准确性。以音频相似度是否达到预设相似度为依据，确定目标音频是否为待处理音频的副本音频，使判定过程具有清晰的量化标准，减少主观因素干扰，确保检测结果的一致性和可靠性。

应理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本申请。

附图说明

图 1 是本申请实施例涉及的实施环境的示意图。

图 2 是本申请实施例示出的音频副本检测方法的第一流程图。

图 3 是本申请实施例示出的音频副本检测方法的第二流程图。

图 4 是本申请实施例示出的提取待处理音频对应的全局音频特征的流程示意图。

图 5 是本申请实施例示出的视频发布平台上用户发表的视频包含的音频分类分布示意图。

图 6 是本申请实施例示出的基于视频发布平台的音频标识处理流程图。

图 7 是本申请实施例示出的针对视频发布平台上发布的视频添加音频标识的原理架构示意图。

图 8 是本申请实施例示出的基于视频发布平台的音频标识处理流程图。

图 9 是本申请实施例示出的视频处理方法的流程图。

图 10 是本申请实施例示出的视频发布平台的前端界面跳转示意图。

图 11 是本申请实施例示出的音频副本检测装置的框图。

图 12 是本申请实施例示出的适于用来实现本申请实施例的电子设备的计算机系统的结构示意图。

需要指出，上述的“第一”、“第二”仅用于区分不同的方案，不代表用于区分方案的优劣程度或在实施过程中的优先级。

具体实施方式

这里将详细地对示例性实施例执行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施

例中所描述的实施方式并不代表与本申请相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

附图中所示的方框图仅仅是功能实体，不一定必须与物理上独立的实体相对应。即，可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络、处理器装置和微控制器装置中的至少之一实现这些功能实体。

附图中所示的流程图仅是示例性说明，不是必须包括所有的内容、以及操作和步骤中的至少之一，也不是必须按所描述的顺序执行。例如，有的操作和步骤中的至少之一还可以分解，而有的操作和步骤中的至少之一可以合并或部分合并，因此实际执行的顺序有可能根据实际情况改变。

在本申请中提及的“多个”是指两个或者两个以上。“至少之一”描述关联对象的关联关系，表示可以存在三种关系，例如，A 和 B 中的至少之一可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。

本申请的说明书和权利要求书及所述附图中的术语“第一”、“第二”、“第三”和“第四”等是用于区别不同对象，而不是用于描述特定顺序。术语“包括”和“具有”以及它们任何变形，意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元，而是可选地还包括没有列出的步骤或单元，或可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。

本申请实施例中，术语“模块”或“单元”是指有预定功能的计算机程序或计算机程序的一部分，并与其他相关部分一起工作以实现预定目标，并且可以通过使用软件、硬件（如处理电路或存储器)或其组合来全部或部分实现。同样的，一个处理器（或多个处理器或存储器)可以用来实现一个或多个模块或单元。此外，每个模块或单元都可以是包含该模块或单元功能的整体模块或单元的一部分。

本申请实施例中相关数据收集处理在实例应用时应该严格根据相关法律法规的要求，获取个人信息主体的知情同意或单独同意，并在法律法规及个人信息主体的授权范围内，开展后续数据使用及处理。

对本申请实施例进行进一步详细说明之前，对本申请实施例中涉及的名词和术语进行说明，本申请实施例中涉及的名词和术语适用于如下的解释。

1）全局音频特征，是指通过对整段音频进行全面、整体的分析提取得到的特征，能够综合反映该音频在整体上的属性，包括但不限于整体的频谱分布趋势、长时程的能量变化规律以及宏观的风格特征等。全局音频特征可通过训练好的特征提取模型对音频的频谱图序列进行处理后获得，是对音频整体内容的抽象表征。

2）音频段特征，是指将音频按照预设的单位时长进行划分，得到若干个连续的音频段后，针对每个音频段单独进行特征提取所得到的特征。音频段特征能够捕捉音频在特定短时间段内的局部特性，如该时间段内的频率成分、能量起伏等细节信息，每个音频段都有其对应的音频段特征。

3）音频段特征序列，是指将音频依据单位时长分割为多个有序的音频段，对每个音频段分别提取音频特征之后，按照这些音频段在原始音频中固有的时间先后顺序，将提取到的音频特征依次排列组合而形成的序列。音频段特征序列完整保留了音频在时间维度上的局部特征变化情况，能体现音频局部特性的时序关联。

4）浅层特征，是由特征提取模型中的浅层特征提取网络生成的特征，浅层特征提取网络对输入的频谱图或者音频段进行初步处理后输出浅层特征。浅层特征更侧重于反

映音频的基础、表层特征，如频谱图中的局部纹理、音频段内的短时频率分布等较为直观的信息，是后续特征处理的基础。

5）深层特征，是在特征提取模型中，由深层特征提取网络对浅层特征进行进一步的整合、抽象和加工处理后得到的特征。深层特征能够穿透音频的表层信息，体现出音频更为抽象、复杂的全局关联性和内在规律，是对音频整体特性更深入的表征。

6）音频相似时长，是对待处理音频和目标音频各自的音频段特征序列进行细致比对，通过分析两个音频段特征序列中特征匹配的情况，确定出两段音频中内容相似的部分，这些相似部分在时间维度上所对应的总时长即为音频相似时长，能够量化反映两个音频内容的相似程度。

首先，如前所述，音频副本检测技术是一种用于识别音频内容是否与已知音频相似或者相同的技术。在相关技术中，主要通过如下方式来实现音频副本检测：

1、基于指纹特征的检测方式：需要提取音频指纹特征来表示指纹内容，通过比较音频指纹之间的相似度，可以判断是否为相同或者相似的音频；

2、基于时域特征的检测方式：需要分析音频的时域特征（例如能量、过零率等），通过计算音频之间的时域特征相似度，来判断是否为相同或者相似的音频；

3、基于频域特征的检测方式：需要分析音频的频域特征（如频谱、谐波等），通过计算音频之间的频域特征相似度，来判断是否为相同或者相似的音频；

4、基于时频域特征的检测方式：需要分析音频的时频域特征（如梅尔频率倒谱系数、声调频率等），通过计算音频之间的时频域特征相似度，来判断是否为相同或者相似的音频；

5、基于深度学习的检测方式：使用深度学习模型（例如卷积神经网络、循环神经网络、长短时记忆网络等）对于音频特征进行编码，并对编码得到的音频特征进行比较，以判断是否为相同或者相似的音频。

在基于深度学习的检测方式中，深度学习模型可以自动学习得到音频的高级特征表示，从而可以提高副本检测的准确性和鲁棒性。由此，本申请实施例所提出的音频副本检测方案，也是采用的基于深度学习的检测方式来判断音频内容是否相同或者相似。

以下将对本申请实施例提出的音频副本检测方案进行详细介绍。

首先参见图 1，图 1 是本申请实施例涉及的实施环境的示意图。该实施环境包括终端 110 和服务器 120，终端 110 和服务器 120 之间通过有线或者无线的方式进行通信。

终端 110 是与用户进行信息交互的设备，用于接收用户输入的信息，比如音频，以及向用户返回信息。例如，终端 110 配置有图形用户界面（Graphical User Interface，GUI），用户可以通过该图形用户界面与终端 110 进行人机交互，比如用户触发图形用户界面以输入信息，也能够从图形用户界面上获得需要的信息。又例如，终端 110 可以配置有键盘、鼠标等设备，用以辅助实现与用户之间的信息交互。

服务器 120 用于为终端 110 与用户之间的信息交互提供数据服务。比如，终端 110 将用户输入的音频上传至服务器 120，使得服务器 120 执行与用户输入的音频相同或者相似的副本音频的检测，并将检测结果返回给终端 110。终端 110 在接收到服务器 120 传输的信息后，将信息返回给用户。

当然，在终端 110 不存在数据服务需求的情况下，终端 110 也可以不与服务器 120 进行数据交互。比如，终端 110 在接收到用户输入的视频后，可以直接进行与用户输入的音频相同或者相似的副本音频的检测，并将检测结果返回给用户。

需要说明的是，终端 110 可以是智能手机、平板、笔记本电脑、计算机、智能家电、智能终端等设备，服务器 120 可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供基础云计算服务的云服务器，本申请实施

例并不对此进行限制。

请参阅图 2，图 2 是本申请实施例示出的音频副本检测方法的第一流程图。该方法可以应用于图 1 所示的实施环境，例如可以由服务器 120 具体执行，也可以由终端 110 具体执行，或者由终端 110 和服务器 120 共同执行。当然，该方法也可以应用于其他的实施环境，并由其他实施环境中的终端或者服务器执行，或者由其他实施环境中的终端和服务器共同执行，本实施例并不进行限制。

如图 2 所示，在一示例性的实施例中，音频副本检测方法包括 S210-S240，详细介绍如下：

S210，提取待处理音频对应的全局音频特征，并提取待处理音频对应的音频段特征序列。

首先说明的是，本实施例揭示的全局音频特征用于表示音频的整体性质，全局音频特征通常需要多层神经网络进行提取得到，因此全局音频特征也可以理解为是深层特征。

本实施例揭示的音频段特征序列是由多个音频段特征按照时间先后顺序排列形成。示例性的，通过将音频划分为多个音频段，然后分别提取各个音频段的音频特征，并按照时间先后顺序对各个音频特征进行排序，从而得到相应的音频段特征序列。

本实施例基于待处理音频对应的全局音频特征，可以实现与待处理音频相似的目标音频的召回；基于待处理音频对应的音频段特征序列，可以实现对召回的目标音频是否与待处理音频相似的进一步校验，由此保证所得到的检测结果具有较高的准确性。

示例的，若待处理音频是一首时长 3 分钟的歌曲，通过多层神经网络提取的全局音频特征可以是一个 128 维的深层特征，综合体现歌曲的整体风格、节奏框架等全局性质；将歌曲按 0.5 秒一段分割后，得到 360 个音频段，提取每个音频段对应的 64 维音频段特征，并且按照音频段在歌曲中的时间顺序，将多个音频段特征排列为音频段特征序列，音频段特征序列包括 360 个 64 维向量的音频段特征。

S220，从音频检索库中检索该全局音频特征对应的近似全局音频特征。

音频检索库中存储了大量的已知音频的全局音频特征。已知音频的全局音频特征与待处理音频的全局音频特征通过同样的特征提取方式提取得到。

示例性的，通过计算待处理音频的全局音频特征与已知音频的全局音频特征之间的相似度，则可以将相似度大于或者等于预设阈值的已知音频的全局音频特征确定为是近似全局音频特征，并将相似度大于或者等于预设阈值的已知音频称为目标音频。又或者，通过将待处理音频和已知音频分别的全局音频特征转换为哈希值，通过比较哈希值来快速判断两个全局音频特征是否相同或者相似。

示例的，还可以将最高的相似度对应的已知音频作为目标音频。针对每个已知音频的全局音频特征，待处理音频的全局音频特征与已知音频的全局音频特征之间的相似度，可以通过以下任意一种算法实现：余弦相似度、欧氏距离、曼哈顿距离和皮尔逊相关系数。以余弦相似度为例，确定待处理音频的全局音频特征与已知音频的全局音频特征的向量积；确定待处理音频的全局音频特征的长度与已知音频的全局音频特征的长度的长度乘积；确定向量积和长度乘积的比值，以作为相似度。例如，待处理音频的全局音频特征为 m，已知音频的全局音频特征为 n，则向量积为 m*n，待处理音频的全局音频特征的长度为|m|，已知音频的全局音频特征的长度为|n|，则长度乘积为|m|*|n|，向量积和长度乘积的比值为（m*n）/（|m|*|n|），以作为相似度。以比较哈希值选择全局音频特征为例，通过哈希函数将待处理音频的全局音频特征映射为第一哈希值，将已知音频的全局音频特征映射为第二哈希值，确定第一哈希值与第二哈希值之间的汉明距离，若汉明距离大于零且小于汉明距离阈值，则确定待处理音频的全局音频特征与已知音频的全局音频特征相似，若汉明距离等于零汉明距离阈值，则确定待处理音频的全局音频特征

与已知音频的全局音频特征相同，若汉明距离大于或等于汉明距离阈值，则确定待处理音频的全局音频特征与已知音频的全局音频特征不相同也不相似。

需要说明的是，音频检索库中，近似全局音频特征的数量可以是一个，也可以是多个，也可以不存在，本实施例并不对检索得到的近似全局音频特征的数量进行限制。在不存在近似全局音频特征的情况下，表示待处理音频相对于音频检索库来说是新的音频。

S230，获取近似全局音频特征所属的目标音频对应的音频段特征序列，并根据待处理音频和目标音频分别对应的音频段特征序列，确定待处理音频与目标音频之间的音频相似度。

在本实施例中，已知音频也对应有全局音频特征和音频段特征序列。已知音频对应的全局音频特征和音频段特征序列的提取方式，与待处理音频对应的全局音频特征和音频段特征序列的提取方式可以相同。

在一些示例性的实施例中，已知音频的全局音频特征和音频段特征序列都可以存储在音频检索库中，并且可以进行关联存储，因此，在从音频检索库中检索到近似全局音频特征之后，可以在音频检索库中查找与该近似全局音频特征关联的音频段特征序列，从而得到该近似全局音频特征所属的目标音频对应的音频段特征序列。

示例性的，在音频检索库中，已知音频的全局音频特征与音频段特征序列可以进行键值对存储，其中全局音频特征作为键，组成音频段特征序列的每一音频段特征分别作为不同的值。

或者，已知音频的全局音频特征和音频段特征序列在音频检索库中可以是分别存储的。比如，音频检索库中存储的各个全局音频特征和各个音频段特征序列分别具有音频标识，相同音频标识对应的音频也应是相同或相似的。

因此在一些示例性实施例中，在检测出待处理视频对应的副本音频后，还将副本音频对应的音频标识作为待处理音频对应的音频标识，并基于待处理音频对应的音频标识，将待处理音频对应的全局音频特征和音频段特征序列存储于音频检索库中。

在另一些示例性实施例中，已知音频的音频段特征序列存储在不同于音频检索库的其他数据库中。比如，每一已知音频均有各自的音频标识，在音频检索库中将已知音频的音频标识与已知音频的全局音频特征关联存储，在其他数据库中将已知音频的音频标识与已知音频的音频段特征序列关联存储。由此，可以根据近似全局音频特征关联的音频标识，从其他数据库中获取该音频标识关联的音频段特征序列，从而得到该近似全局音频特征所属的目标音频对应的音频段特征序列。

示例性的，在其他数据库中，已知音频的音频标识与音频段特征序列可以进行键值对存储，其中音频标识作为键，组成音频段特征序列的每一音频段特征分别作为不同的值。由此，其他数据库可以是键值数据库。

还需要说明的是，全局音频特征和音频段特征序列在数据库中的存储可以具有指定时长的存储时限，也即，数据库只会对全局音频特征和音频段特征序列保留指定时长。因此，在一些示例性的实施例中，在存储全局音频特征和音频段特征序列时，当前存储的全局音频特征和音频段特征序列自动获得指定时长的存储时限，还可以将副本音频对应的全局音频特征和音频段特征序列的剩余存储时长更新为指定时长，由此保证数据库随着音频的新增而更新迭代。

本实施例基于待处理音频和目标音频分别对应的音频段特征序列，通过进行音频段特征序列之间的对比，可以理解为是通过对比音频段特征序列中的音频段特征，来确定出待处理音频和目标音频中的相同音频段，从而确定待处理音频与目标音频之间的音频相似度。

示例性的，可以通过对待处理音频对应的音频段特征序列和目标音频对应的音频段

特征序列进行音频段特征比对，获得待处理音频与目标音频之间的音频相似时长，并根据待处理音频和目标音频分别对应的音频时长，以及音频相似时长，来确定出待处理音频与目标音频之间的音频相似度。

例如，可以先从待处理音频的时长和目标音频的时长中确定出较小的时长，然后计算音频相似时长与该较小时长之间的比值，以得到待处理音频与目标音频之间的音频相似度。又例如，也可以根据待处理音频的时长和目标音频的时长中的较大时长来计算出待处理音频与目标音频之间的音频相似度，本实施例对此并不进行限制。

为便于理解，举例来说，假设待处理视频的时长为 110 秒，目标视频的时长为 120 秒，通过二者对应的音频段特征序列之间的对比，确定待处理视频从第 10 秒到第 110 秒，与目标视频的第 20 秒到第 120 秒，这 100 秒的音频是相似的，则可得到音频相似时长为 100 秒，音频相似度则为：100 秒/min（110 秒，120 秒）=91%。

在一些实施例中，对待处理音频的音频段特征序列和目标音频的音频段特征序列进行时间轴对齐，确保两段音频段特征序列的时间粒度一致；采用向量相似度度量方法，逐一计算待处理音频的音频段特征与目标音频中对应在相同时间段的音频段特征的相似度，设定特征匹配阈值，若待处理音频的音频段特征和目标音频的音频段特征的相似度大于或等于特征匹配阈值，则确定该时间段的待处理音频和目标音频相似；将连续的相似音频段合并为音频相似区间，避免因单段特征波动导致的碎片化判断。将待处理音频中的音频相似区间转换为第一实际时间区间，将目标音频中的音频相似区间转换为第二实际时间区间；将第一实际时间区间与第二实际时间区间中重叠部分的时长作为音频相似时长。

示例的，待处理音频的音频段特征序列包括 220 个特征向量（对应 110 秒，每 0.5 秒一段），目标音频的序列包括 240 个特征向量（对应 120 秒）。逐段计算向量相似度后，发现待处理音频的第 20 至 220 段（对应 10 至 110 秒）与目标音频的第 40 至 240 段（对应 20 至 120 秒）中，每个音频段的音频段特征的相似度均大于或等于 0.85，且这些相似音频段连续无间断，则将这些音频段合并为一个音频相似区间。待处理音频的第一实际时间区间为 10 至 110 秒（时长 100 秒），目标音频的第二实际时间区间为 20 至 120 秒（时长 100 秒），两者重叠区间为 20 至 110 秒，重叠区间的时长为 100 秒，也即音频相似时长为 100 秒。

S240，若该音频相似度达到预设相似度以上，则将目标音频确定为待处理音频对应的副本音频。

预设相似度可以根据实际的应用需求进行设置。比如，在一些对于副本检测准确性要求较高的应用场景中，预设相似度可以设置为较大的值，比如 90%、85%、80%等。而在一些对于副本检测准确性要求较低的应用场景中，预设相似度可以设置为稍低的值，比如 70%、75%等，在此不做限制。

举例来说，假设在 S220 中检索得到两个近似全局音频特征，也即相应确定两个目标音频，然后基于音频段特征序列进行进一步比对，若得到其中一个目标音频的音频段特征序列与待处理音频的音频段特征序列之间的音频相似度较高，例如高于预设相似度，另一个目标音频的音频段特征序列与待处理音频的音频段特征序列之间的音频相似度较低，例如低于预设相似度，则确定音频相似度高于预设相似度的目标音频为待处理音频对应的副本音频。由此可以得到，本实施例先通过全局音频特征来实现较粗粒度的检测，然后再通过音频段特征序列来实现更细粒度的检测，保证了音频副本检测的准确性。

在一些实施例中，收集样本数据，样本数据包括副本音频对（如同一首歌的不同版本、经剪辑的同源音频）和非副本音频对（如不同歌曲、无关联的音频片段），计算每个音频对中两个音频的音频相似度；基于多个音频对的音频相似度，绘制相似度分布曲

线，观察副本音频对与非副本音频对的相似度阈值临界点。例如，通过接受者操作特性曲线（receiver operating characteristic curve，ROC）找到假阳性率（将非副本误判为副本） 和假阴性率（将副本误判为非副本）平衡的阈值点，或设定目标指标（如准确率大于或等于 95%）对应的最小阈值，以作为预设相似度。

示例的，收集 1000 个副本音频对和 1000 个非副本音频对，计算每个音频对的相似度，其中，98%的副本音频对的相似度大于或等于 82%，99%的非副本音频对的相似度小于或等于 78%，通过 ROC 曲线分析，当预设相似度设为 80%时，假阳性率和假阴性率均低于 3%，因此将预设相似度确定为 80%。

请继续参阅图 3，图 3 是本申请实施例示出的音频副本检测方法的第二流程图。

如图 3 所示，S210 中提取音频段特征序列的过程包括如下 S310-S320 内容:

S310，基于预设的第一单位时长对待处理音频进行划分处理，得到待处理音频包含的多个第一音频段;

在一些实施例中，从待处理音频的起始时刻开始，按照第一单位时长对待处理音频进行连续分割，例如，以起始时刻为第一个第一音频段的起点，持续第一单位时长后为第一个第一音频段的终点；紧接着以第一个第一音频段的终点为第二个第一音频段的起点，继续分割，直至覆盖待处理音频的全部时长。若待处理音频的总时长不是第一单位时长的整数倍，最后一个第一音频段的时长为剩余时长。

示例的，待处理音频的总时长为 10.3 秒，预设的第一单位时长为 1 秒。从 0 秒开始分割，得到 10 个完整的第一音频段（分别对应 0 至 1 秒、1 至 2 秒、……、9 至 10 秒），以及第 11 个第一音频段（对应 10-10.3 秒），共 11 个第一音频段。

S320，对各个第一音频段分别提取音频特征，并按照时间先后顺序对各个第一音频段的音频特征进行排序，得到待处理音频对应的音频段特征序列。

可以理解，在 S310 中，第一单位时长通常是一个较小的数值，比如 1 秒，基于第一单位时长对待处理音频进行划分处理，以得到多个第一单位时长的第一音频段。

在 S320 中，通过对各个第一音频段分别提取音频特征，然后按照时间先后顺序对各个第一音频段的音频特征进行排序，从而得到待处理音频对应的音频段特征序列。比如，假设待处理音频的总时长为 100 秒，第一单位时长为 1 秒，待处理音频对应的音频段特征序列则是由 100 个第一音频段的音频特征按照时间先后顺序排列得到的。由此可以看出，音频段特征序列能够表征出更细粒度的音频内容。

本申请实施例通过固定的第一单位时长划分音频段，再按时间顺序排列音频段的音频段特征，确保音频段特征序列严格对应原始音频的时间流向，为后续与目标音频段特征序列比对提供准确的时序基础。将待处理音频拆分为多个第一音频段并分别提取音频段特征，能细致捕捉不同时间段的局部音频特性，避免因全局特征平滑化导致的局部细节丢失，提升对局部音频的检测敏感性。基于预设第一单位时长的划分规则固定，使得不同待处理音频段特征序列具有统一的时间粒度，确保后续跨音频比对的公平性和结果的可重复验证。

继续参见图 3，S210 中提取待处理音频对应的全局音频特征的过程还包括如下 S330-S350 所示内容:

S330，基于预设的第二单位时长对待处理音频进行划分处理，得到待处理音频包含的多个第二音频段;

在一些实施例中，从待处理音频的起始时刻开始，按照第二单位时长对待处理音频进行连续分割，例如，以起始时刻为第一个第二音频段的起点，持续第二单位时长后为第一个第二音频段的终点；紧接着以第一个第二音频段的终点为第二个第二音频段的起点，继续分割，直至覆盖待处理音频的全部时长。若待处理音频的总时长不是第二单位

时长的整数倍，最后一个第二音频段的时长为剩余时长。

示例的，待处理音频的总时长为 10.3 秒，预设的第一单位时长为 1 秒，预设的第二单位时长为 2 秒（大于第一单位时长）。从 0 秒开始分割，得到 5 个完整的第二音频段（分别对应 0 至 2 秒、2 至 4 秒、4 至 6 秒、6 至 8 秒、8 至 10 秒），以及第 6 个第二音频段（对应 10 至 10.3 秒），共 6 个第二音频段。

S340，对各个第二音频段分别进行频谱图转换，得到待处理音频对应的频谱图序列；

这里，若第二音频段存在静音部分，可进行静音切除，保留有效音频信号；对每个第二音频段进行统一的采样率调整，确保不同第二音频段的时域信号在时间维度上的分辨率一致；对预处理后的每个第二音频段，执行以下处理：将第二音频段的时域信号按固定时间窗口划分为多个帧，相邻帧之间设置重叠，以减少帧间信息丢失；对每个帧信号施加窗函数（如汉宁窗），降低帧边缘的频谱泄露；对加窗后的每一帧进行短时傅里叶变换，将时域信号转换为频域信号，得到每个帧的频谱；将同一第二音频段中所有帧的频谱按时间顺序排列，形成二维矩阵（横轴为时间帧，纵轴为频率点，矩阵元素值为对应频率的能量），该二维矩阵即为该第二音频段对应的频谱图。

S350，将频谱图序列输入训练好的特征提取模型中，得到训练好的特征提取模型输出的全局音频特征。

可以理解，由于待处理音频的总时长通常较大，如果直接针对待处理音频进行全局特征的提取，容易造成深度学习网络的负载太大而出现崩溃，因此 S330 中基于第二单位时长对待处理音频进行划分处理，将待处理音频划分为多个第二音频段，后续在第二音频段的基础上提取全局音频特征，由此在一定程度上降低深度学习网络的负载。

由此，基于不同的出发点，第二单位时长通常大于或者等于第一单位时长。通常情况下，第二单位时长大于第一单位时长，例如第二单位时长可以是 10 秒。而在音频段时长不足第二单位时长的情况下，可以通过对音频段进行时长补全的方式来得到符合第二单位时长的第二音频段。

在 S340 中，通过对各个第二音频段进行频谱图转换，以将音频时序的波形转换为梅尔频谱图，由此得到待处理音频对应的频谱图序列。在 S350 中，将待处理音频对应的频谱图序列输入训练好的特征提取模型中，则可以得到训练好的特征提取模型输出的全局音频特征。

请参见图 4，图 4 是本申请实施例示出的提取待处理音频对应的全局音频特征的流程示意图。可以看出，训练好的特征提取模型包括级联的浅层特征提取网络和深层特征提取网络，待处理音频对应的频谱图序列输入至训练好的特征提取网络，先通过浅层特征提取网络提取频谱图序列中的各个频率图分别对应的浅层特征，相应得到待处理音频对应的浅层特征序列，然后通过深层特征提取网络提取各个浅层特征对应的深层特征，最后根据各个深层特征计算出特征平均值，从而将计算得到的特征平均值作为待处理音频对应的全局音频特征。其中，浅层特征提取网络在训练好的特征提取模型中的次序小于深层特征提取网络在训练好的特征提取模型中的次序，例如，浅层特征由特征提取模型靠前的网络层（如初始的卷积层、输入层附近的处理单元）输出，深层特征则由特征提取模型靠后的网络层（如多个卷积层、池化层堆叠后的全连接层或输出层附近的处理单元）输出。

为便于理解，举例来说，假设第二单位时长为 10 秒，每个第二音频段经由训练好的特征提取模型提取出 768 维的特征向量，因此每个第二音频段对应的音频特征可以表示为（batch，768），其中 batch 为待处理音频的时长与第二单位时长之间的商，然后对 batch 维度进行平均值计算，即可得到待处理音频对应的全局音频特征。

例如，浅层特征提取网络可以包括多层 CNN（Convolutional Neural Networks，卷积

神经网络），以利用卷积神经网络的平移不变性来学习浅层特征；深层特征提取网络可以包括 TransFormer（是一种用于序列到序列的深度学习模型架构）网络，以利用 TransFormer 网络的长时记忆能力来学习深层特征。但需要说明的是，本实施例并不对浅层特征提取网络和深层特征提取网络的具体网络结构进行限制，任何可以实现上述功能的浅层特征提取网络和深层特征提取网络的组合均可以用来作为特征提取模型的结构实现。

本申请实施例通过浅层网络聚焦频谱图的基础视觉特征，适合作为细粒度的音频段特征；通过深层网络对浅层特征进一步整合抽象，捕捉跨段的全局关联性，生成的深层特征可以代表音频整体属性，这种分层提取机制，让全局音频特征与音频段特征序列分别对应不同抽象层级，满足不同粒度的比对需求。将深层特征的平均值作为全局音频特征，能综合多个第二音频段的深层特征信息，避免单一第二音频段特征的片面性，更全面反映整体特性，提升全局音频特征的代表性。

基于如上示例的特征提取模型的网络结构，在一示例性的实施例中，S210 可以将待处理音频对应的第一音频段序列输入训练好的特征提取模型，将浅层特征提取网络相应输出的浅层特征序列作为待处理音频对应的音频段特征序列。也即，待处理音频对应的音频段特征序列中包含的音频段特征是浅层特征，浅层特征不具有语义性，能够更准确地表达音频框架，因此在使用全局音频特征实现相似全局音频特征的召回后，通过浅层特征序列来进一步进行特征相似性的校验，能够进一步提升音频副本检测的准确性。

还需要说明的是，训练好的特征提取模型是利用预先收集的训练数据集对初始的特征提取模型进行预训练得到的。预训练的过程即是自监督学习的过程，自监督学习的基本思想是通过对样本数据进行某种变换（例如旋转、遮挡、颜色变换等），使得特征提取模型可以从这些变换后的数据中学习到有用的特征表示。

在一些示例性的实施例中，可以分别生成重建学习任务和对比学习任务中的至少一种训练任务，然后基于训练数据集分别执行该至少一种训练任务，对初始的特征提取模型进行预训练，得到训练好的特征提取模型。

重建学习任务指示通过计算平均绝对误差（Mean Absolute Error，MAE）损失值来进行模型参数优化。重建学习任务对应的损失函数的计算公式可以表示如公式（1）：

$$loss(MAE) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

其中，n表示一个训练批次中的样本数量，$\hat{y}_i$表示初始的特征提取模型输出的全局音频特征，y_i表示真实的全局音频特征，$loss(MAE)$表征重建学习任务对应的损失函数，Σ用于对数据进行求和。

对比学习任务指示通过计算噪声对比估计损失值来进行模型参数优化。对比学习任务的基本思想是通过对比正样本和负样本来优化模型，模型的目标则是最大化正样本与目标之间的相似度，同时最小化负样本与目标之间的相似度。 对比学习任务对应的损失函数可以表示为 loss(InfoNCE)，计算公式如公式（2）：

$$loss(InfoNCE) = -\frac{1}{n}\sum_{i=1}^{n} log\frac{A_+}{A_+ + B_-} \quad (2)$$

其中，n表示一个训练批次中的样本数量，A_+表示正样本对的相似度得分，B_-表示所有负样本对的相似度得分之和，$loss(InfoNCE)$表征对比学习任务对应的损失函数，Σ用于对数据进行求和，log用于进行对数运算。

在训练过程中，可以通过梯度缩减、特征罚项中的至少一种方式来提升特征提取模型的训练稳定性。可以理解，梯度缩减是通过限制梯度的大小来避免模型参数在更新时发生极端变化，特征罚项是通过在损失函数中添加一个与模型复杂度相关的项，以使得模型在训练数据上表现良好的同时，也保持一定的泛化能力。

另外还需要说明的是，在一些示例性实施例中，可以通过分布式数据爬取的方式来

收集训练数据，由此得到多元异构的训练数据集，进而可以使用大规模的训练数据集来对初始的特征提取模型执行预训练。比如，训练数据来源于不同类型的音频业务，与这些音频业务有关的数据是进行分布式存储的，通过分布式爬取这些音频业务的音频数据，即可得到由不同类型的音频数据所构成的训练数据集。

本申请实施例通过重建学习任务促使特征提取模型捕捉频谱图的细节信息，提升对音频局部特征的还原能力，通过对比学习任务推动特征提取模型区分相似与非相似音频的特征差异，增强对全局语义的辨别能力，两种任务结合使得特征提取模型在复杂音频场景中保持稳定的特征提取性能，提高泛化能力。基于训练数据集执行重建学习任务，通过量化平均绝对误差直接反馈特征提取的精准度，对比学习任务通过噪声对比估计快速强化特征提取模型对关键特征的敏感度，避免无目标的参数调整，缩短训练周期并提升模型收敛质量。

在一个示例性的应用场景中，本申请实施例提出的音频副本检测方法应用于视频发布平台。视频发布平台可以理解为是 UGC（User Generated Content，用户生成内容）平台，用户可以在视频发布平台上发表自己制作的视频，也可以在视频发布平台上浏览其他用户发表的视频。

用户在制作待发布的视频作品时，通常会添加背景音乐、音效、解说、独白等方式来辅助表现视频内容，以赋予视频场景感，带动观众的情绪氛围。图 5 是本申请实施例示出的视频发布平台上用户发表的视频包含的音频分类分布示意图，可以看出，使用纯音乐发布视频的用户的占比并不大，多数用户习惯于使用多种音频的组合来发表视频。在此应用场景下，如果要进行音频副本检测，仅考虑视频作品包含的背景音乐是不准确的，因此本申请的实施例还提出基于视频的原声来实现音频副本检测。

视频发布平台中发布的视频中包含的音频对应的全局音频特征存储于音频检索库中，并且在音频检索库中，各个全局音频特征分别具有音频标识。这也可以理解为，相同的音频对应于相同的音频标识，不同的音频对应不同的音频标识。

请参见图 6，图 6 是本申请实施例示出的基于视频发布平台的音频标识处理流程图。可以看出，针对视频发布平台上发布的视频，首先提取出视频包含的音频信息作为待处理音频，然后提取待处理音频对应的全局音频特征和音频段特征序列，再基于音频检索库检索全局音频特征对应的近似全局音频特征。若检索到近似全局音频特征，则表示检索到了相同音频特征，将此近似全局音频特征对应的音频标识添加为待处理音频的音频标识。若未检索到近似全局音频特征，则基于音频标识列表中最新的音频标识，生成待处理音频对应的音频标识。

音频标识列表记录了视频发布平台上发布的视频所包含的音频对应的音频标识，因此，本实施例基于音频标识列表中最新的音频标识，生成待处理音频对应的音频标识，可以保证所生成的音频标识是音频标识列表中不存在的新的音频标识。比如，可以根据预设的数值更新规则，如数值每次增加预设值（例如 1），对音频标识列表中最新的音频标识执行数值更新处理，将得到的数值作为待处理音频的音频标识。需要说明的是，本实施例并不对音频标识的具体更新方式进行限制，比如每次增加的数值可以为除 1 以外的其他值，也可以通过每次执行数值减小的方式来生成新的音频标识。

示例的，数值更新规则还包括：对音频标识列表中最新的音频标识进行哈希运算，将运算结果作为更新后的数值，也即待处理音频对应的音频标识，或者，将运算结果从头部开始的预设长度的数值，作为待处理音频对应的音频标识。

本申请实施例当音频相似度达到预设相似度以上，将副本音频的音频标识赋予待处理音频，使内容相似的音频共用同一音频标识，减少数据冗余；若音频相似度低于预设相似度，基于音频标识列表中最新的音频标识生成待处理音频对应的音频标识，保证待

处理音频的音频标识与已有标识进行区分，实现音频标识与音频内容的准确对应。

请参阅图 7，图 7 是本申请实施例示出的针对视频发布平台上发布的视频添加音频标识的原理架构示意图。针对视频发布平台上发布的视频，分别提取全局音频特征作为视频级别特征，提取音频段特征序列作为帧级别特征，然后基于音频检索库执行相似召回，以基于视频级别特征召回相似音频，并且基于键值数据库执行重复校验，以基于帧级别特征实现副本音频的重复校验，从而根据重复校验的结果来为视频添加音频标识。需要理解的是，图 7 示意的音频检索库是用于存储全局音频特征，键值数据库是用于存储音频段特征序列。

另外需要说明的是，如图 6 所示，音频检索库不仅用于存储全局音频特征和音频段特征序列，还可以存储音频标签信息，如音频的曲风、语种、情感等标签类型，在此不进行限制。这些音频标签可以在一定程度上对所属的视频进行区分，以满足某些视频处理场景。比如，可以通过音频检索库中存储的音频标签来快速筛选出具有某类风格的音频的视频。

在视频发布平台的应用场景下，如图 8 所示，音频副本检测方法还可以包括 S810-S820，详细介绍如下：

S810，检测视频发布平台中的待推荐视频，并获取待推荐视频包含的音频对应的音频标识。

本实施例提供的技术方案是应用于视频发布平台中的视频推荐场景。可以基于具体推荐需求来检测视频发布平台中的待推荐视频。比如，具体推荐需求为针对用户新发布的视频执行视频推荐，则将视频发布平台中新发布的视频检测为待推荐视频。又比如，具体推荐需求为针对具有特定标签的用户所发布的视频执行视频推荐，特定标签例如表示用户在视频发布平台中进行了付费操作和充值操作中的至少之一，则将视频发布平台中具有此特定标签的用户发布的视频作为待推荐视频。

待推荐视频包含的音频对应的音频标识，也即是按照如图 6 所示例的流程得到的，在此不再对此流程进行赘述。

S820，基于待推荐视频对应的音频标识，确定待推荐视频对应的候选视频，并基于候选视频执行待推荐视频的推荐处理；其中，候选视频包含的音频为待推荐视频包含的音频对应的副本音频。

由于相同或相似的音频具有相同的音频标识，因此在本实施例中，首先获取音频标识与待推荐视频所包含音频的音频标识相同的副本音频，然后将包含副本音频的视频获取为候选视频，从而基于获取到的候选视频对待推荐视频进行推荐处理。

示例性的，可以通过获取候选视频的观众，并向这些观众推荐待推荐视频。或者，基于这些观众针对候选视频的观看数据，比如观看时长、评价信息等，进一步筛选出对候选视频感兴趣的观众，然后向筛选出的观众推荐待推荐视频。由此，可以实现待推荐视频的精准推荐。

又或者，可以将待推荐视频与候选视频进行打包推荐。比如，将待推荐视频与候选视频作为一个视频包，当任一观众浏览到视频包中的任一视频时，可以按照预设规则向该观众推荐视频包中的其他视频。预设规则例如是观看完当前视频后立即推荐播放视频包中的其他视频，或者是观看完当前视频后，间隔预设数量的视频后再推荐播放视频包中的其他视频，在此不进行限制。

本实施例并不对具体的视频推荐方式进行限制，但需要说明的是，本实施例所进行的视频推荐依赖于对视频所包含音频进行的副本音频识别。在从视频中提取出音频后，即可对提取出的音频进行音频副本检测，由于从视频中提取出的音频是原声音频，原声音频也即是包含了视频中所有的音频信息，比如背景音乐、解说、独白等，这使得本实

施例所进行的视频推荐考虑了视频中的所有音频信息，由此保证具有较好的视频推荐效果。

本申请实施例基于待推荐视频对应的音频标识，获取包含相同音频标识的副本音频的候选视频，进而针对候选视频的观众推荐待推荐视频，利用相同或相似音频的关联关系，确保推荐对象与待推荐视频的匹配度，提升推荐精准性。支持向候选视频的观众直接推荐、基于观看数据筛选后推荐、将待推荐视频与候选视频打包推荐等多种方式，且打包推荐的预设规则可灵活设定，适应不同场景下的推荐需求，增强推荐策略的多样性。

在视频发布平台的应用场景下，采用本申请实施例提供的音频副本检测方案还可以提高视频发布平台中的作品发表率。在另一示例性的实施例中，如图 9 所示，一种由用户终端执行的视频处理方法包括如下步骤：

S910，在视频播放页面进行视频播放；

这里，视频播放页面包括多个视频，响应于针对任一视频的选择操作，在视频页面播放被选择的视频。或者，响应于针对视频播放页面的打开操作，播放视频平台推荐的视频。视频发布平台会加载视频的音视频数据，在视频播放页面中初始化播放器组件，按照视频的原始帧率、分辨率等参数进行解码播放，同时实时渲染视频画面并输出音频，确保用户能正常观看和收听视频内容。

S920，若检测到视频播放页面中设置的副本音频搜索入口被触发，则显示候选视频展示页面；候选视频展示页面显示有多个候选视频，候选视频包含的音频为视频播放页面中播放的视频所包含音频对应的副本音频；

S930，若检测到候选视频展示页面中设置的视频发布入口被触发，则显示视频编辑页面。

需要说明的是，视频发布平台具有前端和后端。前端是指用户直接与之交互的部分，主要负责将后端提供的数据以图形化界面的形式展示给用户，通常是通过网页浏览器展示，并处理用户与界面之间的交互。后端负责处理前端发送的请求、执行相应的逻辑（如访问数据库、进行数据处理等），并将结果返回给前端显示。

在此以图 10 中示意出的视频发布平台的前端界面跳转示意图作为示例，来对 S910-S930 所示的过程进行解释。

用户终端首先显示视频播放页面，并在视频播放页面进行视频播放。视频播放页面中设置有副本音频搜索入口，比如图 10 所示的视频播放页面中的“找同款”按钮也即为副本音频搜索入口。当用户终端检测到该副本音频搜索入口被触发，跳转显示候选视频展示页面。

候选视频展示页面显示有多个候选视频，候选视频包含的音频为视频播放页面中播放的视频所包含音频对应的副本音频，也可以理解为，候选视频对应的音频标识与视频播放页面中播放的视频对应的音频标识是相同的。其中，对视频所包含的音频对应的音频标识的识别，是通过前述实施例中记载的有关流程来实现的。

可以理解的是，候选视频展示页面中显示的候选视频由于使用了相同或者近似的音频，因此候选视频之间可能具有共同内容属性，比如都是搞笑类视频，都是电影解说类视频等，因此，通过在候选视频展示页面中展示相关的候选视频，在一定程度上可以引导用户使用相同音频创作并发布视频作品，因此可以提高视频发布平台中的作品发表率。

另外还需要说明的是，如图 10 所示，候选视频展示页面还可以对视频封面进行显示，比如，候选视频展示页面上显示的多个候选视频为视频封面，该视频封面可以是视频历史发布时由用户设置的封面内容，也可以是直接从候选视频的视频内容中确定出的封面内容，本实施例并不对候选视频的视频封面的获取方式进行限制。

候选视频展示页面中设置有视频发布入口，比如图 10 示出的“发视频”按钮即为

视频发布入口。当用户终端检测到视频发布入口被触发，则跳转显示视频制作页面，用户可以在视频制作页面进行待发布视频的制作操作，本实施例并不对视频制作的具体过程进行限制。但需要说明的是，视频制作页面中可以将此相同音频默认作为添加的背景音乐，以便于用户进行视频制作。

由于可以得到，本实施例将视频发布平台的前端界面中的页面显示与音频副本检测技术相结合，十分便于引导用户在观看视频的过程中，基于相同的音频来制作并发布视频，由此可以提高视频发布平台中的作品发表率，从而可以提高视频发布平台在同类平台产品中的竞争力。

本申请实施例通过副本音频搜索入口，用户可快速找到包含相似音频的候选视频，拓展内容浏览范围；视频发布入口则降低了用户发布关联内容的门槛，提升平台互动性。基于副本音频的关联关系，候选视频展示页面将内容相关的视频集中呈现，使具有相同或相似音频的视频形成内容集群，方便用户高效获取同类信息。

图 11 是本申请实施例示出的音频副本检测装置的框图。如图 11 所示，该示例性的音频副本检测装置包括：

特征提取模块 1110，配置为提取待处理音频对应的全局音频特征，并提取待处理音频对应的音频段特征序列；

特征检索模块 1120，配置为从音频检索库中检索全局音频特征对应的近似全局音频特征，并获取近似全局音频特征所属的目标音频对应的音频段特征序列；

相似度确定模块 1130，配置为根据待处理音频对应的音频段特征序列和目标音频对应的音频段特征序列，确定待处理音频与目标音频之间的音频相似度；

副本判定模块 1140，配置为若音频相似度达到预设相似度以上，则将目标音频确定为待处理音频对应的副本音频。

在另一示例性的实施例中，特征提取模块 1110 包括音频段特征提取单元，音频段特征提取单元配置为：基于预设的第一单位时长对待处理音频进行划分处理，得到待处理音频包含的多个第一音频段；对各个第一音频段分别提取音频特征，并按照时间先后顺序对各个第一音频段的音频特征进行排序，得到待处理音频对应的音频段特征序列。

在另一示例性的实施例中，特征提取模块 1110 包括全局特征提取单元，全局特征提取单元配置为：基于预设的第二单位时长对待处理音频进行划分处理，得到待处理音频包含的多个第二音频段；第二单位时长大于或者等于第一单位时长；对各个第二音频段分别进行频谱图转换，得到待处理音频对应的频谱图序列；将频谱图序列输入训练好的特征提取模型中，得到训练好的特征提取模型输出的全局音频特征。

在另一示例性的实施例中，训练好的特征提取模型是利用训练数据集对初始的特征提取模型进行预训练得到；音频副本检测装置还包括模型训练模块，模型训练模块配置为：分别生成重建学习任务和对比学习任务中的至少一种训练任务；重建学习任务指示通过计算平均绝对误差损失值来进行模型参数优化，对比学习任务指示通过计算噪声对比估计损失值来进行模型参数优化；基于训练数据集分别执行至少一种训练任务，对初始的特征提取模型进行预训练，得到训练好的特征提取模型。

在另一示例性的实施例中，训练好的特征提取模型包括级联的浅层特征提取网络和深层特征提取网络；全局特征提取单元进一步配置为：通过浅层特征提取网络提取频谱图序列中各个频谱图分别对应的浅层特征；通过深层特征提取网络提取各个浅层特征对应的深层特征；根据深层特征计算特征平均值，将特征平均值作为待处理音频对应的全局音频特征。

在另一示例性的实施例中，音频段特征提取单元进一步配置为：按照时间先后顺序将各个第一音频段输入训练好的特征提取模型；将浅层特征提取网络输出的浅层特征序

列作为待处理音频对应的音频段特征序列。

在另一示例性的实施例中，相似度确定模块 1130 进一步配置为：对待处理音频对应的音频段特征序列和目标音频对应的音频段特征序列进行音频段特征比对，获得待处理音频与目标音频之间的音频相似时长；根据待处理音频和目标音频分别对应的音频时长，以及音频相似时长，确定待处理音频与目标音频之间的音频相似度。

在另一示例性的实施例中，音频检索库中存储的各个全局音频特征分别具有音频标识；音频副本检测装置还包括音频存储模块，音频存储模块配置为：将副本音频对应的音频标识作为待处理音频对应的音频标识；基于待处理音频对应的音频标识，将待处理音频对应的全局音频特征和音频段特征序列存储于音频检索库中。

在另一示例性的实施例中，音频检索库中存储的全局音频特征和音频段特征序列具有指定时长的存储时限；音频存储模块进一步配置为：将副本音频对应的全局音频特征和音频段特征序列的剩余存储时长更新为指定时长。

在另一示例性的实施例中，音频存储模块进一步配置为：若音频相似度低于预设相似度，则基于音频标识列表中最新的音频标识，生成待处理音频对应的音频标识。

在另一示例性的实施例中，音频副本检测装置还包括：

视频检测模块，配置为检测视频发布平台中的待推荐视频，并获取待推荐视频包含的音频对应的音频标识；

视频推荐模块，配置为基于待推荐视频对应的音频标识，确定待推荐视频对应的候选视频，并基于候选视频执行待推荐视频的推荐处理；其中，候选视频包含的音频为待推荐视频包含的音频对应的副本音频。

需要说明的是，上述实施例所提供的装置与上述实施例所提供的方法属于同一构思，其中各个模块和单元执行操作的具体方式已经在方法实施例中进行了详细描述，此处不再赘述。上述实施例所提供的音频副本检测装置在实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将装置的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能，本处也不对此进行限制。

本申请的实施例还提供了一种电子设备，包括：一个或多个处理器；存储器，用于存储一个或多个计算机程序，当所述一个或多个计算机程序被所述一个或多个处理器执行时，使得所述电子设备实现上述各个实施例中提供的音频副本检测方法。

图 12 是本申请实施例示出的适于用来实现本申请实施例的电子设备的计算机系统的结构示意图。需要说明的是，该电子设备可以是图 1 所示实施环境中的终端 110 或者服务器 120，也可以是其他实施环境中的终端或者服务器，在此不进行限制。还需要说明的是，图 12 示出的电子设备的计算机系统 1200 仅是一个示例，不应对本申请实施例的功能和使用范围带来任何限制。

如图 12 所示，计算机系统 1200 包括中央处理单元（Central Processing Unit，CPU)1201，其可以根据存储在只读存储器（Read-Only Memory，ROM)1202 中的计算机程序或者从储存部分 1208 加载到随机访问存储器（Random Access Memory，RAM)1203 中的计算机程序而执行各种适当的动作和处理，例如执行上述实施例中所述的方法。在 RAM 1203 中，还存储有系统操作所需的各种计算机程序和数据。CPU 1201、ROM1202 以及 RAM 1203 通过总线 1204 彼此相连。输入或输出（Input /Output，I/O)接口 1205 也连接至总线 1204。

以下部件连接至 I/O 接口 1205：包括键盘、鼠标等的输入部分 1206；包括诸如阴极射线管（Cathode Ray Tube，CRT)、液晶显示器（Liquid Crystal Display，LCD)等以及扬声器等的输出部分 1207；包括硬盘等的储存部分 1208；以及包括诸如 LAN(Local Area Network，局域网)卡、调制解调器等的网络接口卡的通信部分 1209。通信部分 1209 经

由诸如因特网的网络执行通信处理。驱动器 1210 也根据需要连接至 I/O 接口 1205。可拆卸介质 1211，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器 1210 上，以便于从其上读出的计算机程序根据需要被安装入储存部分 1208。

特别地，根据本申请的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本申请的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的计算机程序。在这样的实施例中，该计算机程序可以执行以下至少一种处理：通过通信部分 1209 从网络上被下载和安装，从可拆卸介质 1211 被安装。在该计算机程序被中央处理单元（CPU)1201 执行时，执行本申请的系统中限定的各种功能。

需要说明的是，本申请实施例所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器（RAM)、只读存储器（ROM)、可擦式可编程只读存储器（Erasable Programmable Read Only Memory，EPROM)、闪存、光纤、便携式紧凑磁盘只读存储器（Compact Disc Read-Only Memory，CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。计算机可读介质上包含的计算机程序可以用任何适当的介质传输，包括但不限于：无线、有线等等，或者上述的任意合适的组合。

附图中的流程图和框图，图示了按照本申请各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。其中，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图或流程图中的每个方框、以及框图或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

描述于本申请实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现，所描述的单元也可以设置在处理器中。其中，这些单元的名称在某种情况下并不构成对该单元本身的限定。

本申请的另一方面还提供了一种计算机可读存储介质，其上存储有计算机程序，该计算机程序被电子设备的处理器执行时实现如前所述的音频副本检测方法。该计算机可读存储介质可以是上述实施例中描述的电子设备中所包含的，也可以是单独存在，而未装配入该电子设备中。

本申请的另一方面还提供了一种计算机程序产品，该计算机程序产品包括计算机程序，该计算机程序存储在计算机可读存储介质中。电子设备的处理器从计算机可读存储介质读取该计算机程序，处理器执行该计算机程序，使得该电子设备执行上述各个实施例中提供的音频副本检测方法。

上述内容，仅为本申请的较佳示例性实施例，并非用于限制本申请的实施方案，本领域普通技术人员根据本申请的主要构思和精神，可以十分方便地进行相应的变通或修改，故本申请的保护范围应以权利要求书所要求的保护范围为准。

可以理解的是，在本申请的具体实施方式中，涉及到视频、音频等相关的数据，当本申请以上实施例运用到具体产品或技术中时，需要获得用户许可或者同意，且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准。

权利要求书

1、一种音频副本检测方法，所述方法由电子设备执行，所述方法包括：

提取待处理音频对应的全局音频特征，并提取所述待处理音频对应的音频段特征序列；

从音频检索库中检索所述全局音频特征对应的近似全局音频特征；

获取所述近似全局音频特征所属的目标音频对应的音频段特征序列，并根据所述待处理音频和所述目标音频分别对应的音频段特征序列，确定所述待处理音频与所述目标音频之间的音频相似度；

若所述音频相似度达到预设相似度以上，则将所述目标音频确定为所述待处理音频对应的副本音频。

2、根据权利要求 1 所述的方法，其中，所述提取待处理音频对应的音频段特征序列，包括：

基于预设的第一单位时长对所述待处理音频进行划分处理，得到所述待处理音频包含的多个第一音频段；

对各个第一音频段分别提取音频特征，并按照时间先后顺序对各个第一音频段的音频特征进行排序，得到所述待处理音频对应的音频段特征序列。

3、根据权利要求 2 所述的方法，其中，所述提取待处理音频对应的全局音频特征，包括：

基于预设的第二单位时长对所述待处理音频进行划分处理，得到所述待处理音频包含的多个第二音频段；所述第二单位时长大于或者等于所述第一单位时长；

对各个第二音频段分别进行频谱图转换，得到所述待处理音频对应的频谱图序列；

将所述频谱图序列输入训练好的特征提取模型中，得到所述训练好的特征提取模型输出的全局音频特征。

4、根据权利要求 3 所述的方法，其中，所述训练好的特征提取模型是利用训练数据集对初始的特征提取模型进行预训练得到；所述方法还包括：

分别生成重建学习任务和对比学习任务中的至少一种训练任务；所述重建学习任务指示通过计算平均绝对误差损失值来进行模型参数优化，所述对比学习任务指示通过计算噪声对比估计损失值来进行模型参数优化；

基于所述训练数据集分别执行所述至少一种训练任务，对所述初始的特征提取模型进行预训练，得到所述训练好的特征提取模型。

5、根据权利要求 3 所述的方法，其中，所述训练好的特征提取模型包括级联的浅层特征提取网络和深层特征提取网络；通过执行如下步骤来得到所述待处理音频对应的全局音频特征：

通过所述浅层特征提取网络提取所述频谱图序列中各个频谱图分别对应的浅层特征；

通过所述深层特征提取网络提取各个浅层特征对应的深层特征；

根据所述深层特征计算特征平均值，将所述特征平均值作为所述待处理音频对应的全局音频特征。

6、根据权利要求 5 所述的方法，其中，所述对各个第一音频段分别提取音频特征，并按照时间先后顺序对各个第一音频段的音频特征进行排序，得到所述待处理音频对应的音频段特征序列，包括：

按照时间先后顺序将各个第一音频段输入所述训练好的特征提取模型；

将所述浅层特征提取网络输出的浅层特征序列作为所述待处理音频对应的音频段特征序列。

7、根据权利要求 3 所述的方法，其中，所述对各个第二音频段分别进行频谱图转换，得到所述待处理音频对应的频谱图序列，包括：

针对每个所述第二音频段，执行以下处理：

将所述第二音频段的时域信号按照预设的时间窗口划分为多个帧；

针对每个所述帧，对所述帧进行傅里叶变化，得到所述帧的频谱；

将多个所述帧分别对应的频谱，按照时间顺序组合为所述第二音频段对应的频谱图；

将多个所述第二音频段分别对应的所述频谱图组合为所述待处理音频对应的频谱图序列。

8、根据权利要求 1-7 任一项所述的方法，其中，所述根据所述待处理音频和所述目标音频分别对应的音频段特征序列，确定所述待处理音频与所述目标音频之间的音频相似度，包括：

对所述待处理音频对应的音频段特征序列和所述目标音频对应的音频段特征序列进行音频段特征比对，获得所述待处理音频与所述目标音频之间的音频相似时长；

根据所述待处理音频和所述目标音频分别对应的音频时长，以及所述音频相似时长，确定所述待处理音频与所述目标音频之间的音频相似度。

9、根据权利要求 1-8 任一项所述的方法，其中，所述音频检索库中存储的各个全局音频特征分别具有音频标识；所述方法还包括：

将所述副本音频对应的音频标识作为所述待处理音频对应的音频标识；

基于所述待处理音频对应的音频标识，将所述待处理音频对应的全局音频特征和音频段特征序列存储于所述音频检索库中。

10、根据权利要求 9 所述的方法，其中，所述音频检索库中存储的全局音频特征和音频段特征序列具有指定时长的存储时限；所述方法还包括：

将所述副本音频对应的全局音频特征和音频段特征序列的剩余存储时长更新为所述指定时长。

11、根据权利要求 9 所述的方法，其中，所述方法还包括：

若所述音频相似度低于所述预设相似度，则基于音频标识列表中最新的音频标识，生成所述待处理音频对应的音频标识。

12、根据权利要求 11 所述的方法，其中，所述基于音频标识列表中最新的音频标识，生成所述待处理音频对应的音频标识，包括：

根据预设的数值更新规则对所述音频标识列表中最新的音频标识进行更新，得到更新后的数值，并将所述更新后的数值作为所述待处理音频对应的音频标识，其中，所述数值更新规则包括以下任意一种：

将所述音频标识列表中最新的音频标识增加第一预设值；

将所述音频标识列表中最新的音频标识减少第二预设值；

对所述音频标识列表中最新的音频标识进行哈希运算。

13、根据权利要求 9 所述的方法，其中，所述方法还包括：

检测视频发布平台中的待推荐视频，并获取所述待推荐视频包含的音频对应的音频标识；

基于所述待推荐视频对应的音频标识，确定所述待推荐视频对应的候选视频，并基于所述候选视频执行所述待推荐视频的推荐处理；其中，所述候选视频包含的音频为所述待推荐视频包含的音频对应的副本音频。

14、根据权利要求 13 所述的方法，其中，所述检测视频发布平台中的待推荐视频，包括:

若针对用户新发布的视频执行视频推荐，则将所述视频发布平台中新发布的视频作为待推荐视频;

若针对具有特定标签的用户所发布的视频执行视频推荐，则将所述视频发布平台中具有所述特定标签的用户发布的视频作为所述待推荐视频，其中，所述特定标签表征用户在所述视频发布平台中进行了付费操作和充值操作中的至少之一。

15、一种音频副本检测装置，所述装置包括:

特征提取模块，配置为提取待处理音频对应的全局音频特征，并提取所述待处理音频对应的音频段特征序列;

特征检索模块，配置为从音频检索库中检索所述全局音频特征对应的近似全局音频特征，并获取所述近似全局音频特征所属的目标音频对应的音频段特征序列;

相似度确定模块，配置为根据所述待处理音频对应的音频段特征序列和所述目标音频对应的音频段特征序列，确定所述待处理音频与所述目标音频之间的音频相似度;

副本判定模块，配置为若所述音频相似度达到预设相似度以上，则将所述目标音频确定为所述待处理音频对应的副本音频。

16、一种电子设备，包括:

一个或多个处理器;

存储器，用于存储一个或多个计算机程序，当所述一个或多个计算机程序被所述一个或多个处理器执行时，使得所述电子设备实现如权利要求 1-14 中任一项所述的方法。

17、一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序被电子设备的处理器执行时，使所述电子设备执行权利要求 1-14 中任一项所述的方法。

18、一种计算机程序产品，包括计算机程序，该计算机程序被电子设备的处理器执行时实现如权利要求 1-14 中任一项所述的方法。