

一种音频检索识别方法及装置

5 相关申请的交叉引用

本申请要求申请号为 201810273699.7、申请日为 2018 年 3 月 29 日的中国专利申请的优先权，该文献的全部内容以引用方式并入本文。

技术领域

10 本公开涉及音频处理技术领域，特别是涉及一种音频检索识别方法及装置。

背景技术

15 音频指纹（或者称为音频特征）以及音频指纹检索在如今的“多媒体信息社会”中具有广泛的应用。音频指纹检索最初被应用到听歌识曲之中，也就是输入一段音频，通过提取和比对该音频的指纹特征，就能识别出对应的歌曲。另外，音频指纹检索也可应用到内容监控之中，比如音频消重、基于检索的语音广告监控、音频版权等。

20 现有的音频检索识别方法存在准确性差、速度慢的问题，这对运算资源和存储资源都会产生巨大消耗。

发明内容

本公开的目的在于提供一种新的音频检索识别方法及装置。

25 本公开的目的是采用以下的技术方案来实现的。依据本公开提出的音频检索识别方法，包括以下步骤：获取待识别音频的音频指纹，其中，所述音频指纹包括用于表示所述待识别音频的内容特征的第一部分以及用于表示所述第一部分的可信程度的第二部分；根据所述音频指纹对所述待识别音频进行识别，得到识别结果。

本公开的目的还可以采用以下的技术措施来进一步实现。

30 前述的音频检索识别方法，其中，所述获取待识别音频的音频指纹包括：将所述待识别音频转换成声谱图；确定所述声谱图中的特征点；在所述声谱图上，为所述特征点确定一个或多个掩模，每个所述掩模包含多个谱区域；确定每个所述谱区域的均值能量；根据所述掩模中的所述多个谱区域的均值能量确定音频指纹比特；判断所述音频指纹比特的可信程度以
35 确定强弱权重比特；根据所述音频指纹比特和所述强弱权重比特确定所述待识别音频的音频指纹。

前述的音频检索识别方法，其中，所述将所述待识别音频转换成声谱图包括：通过短时傅里叶变换将所述待识别音频转换成时间-频率的二维声

谱图，所述声谱图中每个点的取值代表所述待识别音频的能量。

前述的音频检索识别方法，其中，所述将所述待识别音频转换成声谱图还包括：对所述声谱图进行梅尔变化。

5 前述的音频检索识别方法，其中，所述特征点为所述声谱图中的固定点。

前述的音频检索识别方法，其中，所述特征点为频率值与预设的多个频率设定值相等的点。

前述的音频检索识别方法，其中，所述特征点为所述声谱图中的能量极大值点，或者，所述特征点为所述声谱图中的能量极小值点。

10 前述的音频检索识别方法，其中，所述掩模所包含的多个所述谱区域是对称分布的。

前述的音频检索识别方法，其中，所述掩模所包含的多个所述谱区域具有相同的频率范围、和/或具有相同的时间范围、和/或以所述特征点为中心而中心对称分布。

15 前述的音频检索识别方法，其中，所述谱区域均值能量为所述谱区域所包含的所有点的能量值的平均值。

前述的音频检索识别方法，其中，所述的根据所述掩模中的所述多个谱区域的均值能量确定音频指纹比特包括：根据一个所述掩模所包含的多个所述谱区域的均值能量的差值确定一个音频指纹比特。

20 前述的音频检索识别方法，其中，所述的判断所述音频指纹比特的可信程度以确定强弱权重比特包括：判断所述差值的绝对值是否达到或超过预设的强弱比特阈值，如果达到或超过所述强弱比特阈值，则将所述音频指纹比特确定为强比特，否则降所述音频指纹比特确定为弱比特；根据所述音频指纹比特是强比特还是弱比特来确定所述强弱权重比特。

25 前述的音频检索识别方法，还包括：将待识别音频按时间分成多段子音频；提取每段所述子音频的所述音频指纹；将提取得到的各个所述子音频的所述音频指纹进行组合，得到所述待识别音频的音频指纹。

30 前述的音频检索识别方法，其中，将所述待识别音频的音频指纹定义为第一音频指纹，所述第一音频指纹包含多个第一音频指纹单体以及与各个所述第一音频指纹单体相对应的第一强弱权重单体，所述第一音频指纹单体包含所述待识别音频的多个所述音频指纹比特，所述第一强弱权重单体包含与所述多个音频指纹比特相对应的多个所述强弱权重比特。

35 前述的音频检索识别方法，其中，所述根据所述音频指纹对所述待识别音频进行识别包括：根据每个单独的所述第一音频指纹单体对多个已知音频进行第一排名，根据所述第一排名的结果，取出前 k 个所述已知音频作为第一候选音频集合，其中 k 为正整数；根据多个顺序排列的所述第一

音频指纹单体对所述第一候选音频集合进行第二排名，根据所述第二排名的结果，取出前 n 个所述第一候选音频作为识别结果，其中 n 为正整数。

前述的音频检索识别方法，还包括：预先获取所述已知音频的音频指纹作为第二音频指纹，所述第二音频指纹包含多个第二音频指纹单体以及
5 与所述第二音频指纹单体相对应的第二强弱权重单体；对所述第二音频指纹进行索引，以预先得到所述已知音频的指纹索引。

前述的音频检索识别方法，其中，在进行所述第一排名和/或进行所述第二排名的过程中，根据所述第一强弱权重单体和/或第二强弱权重单体，对所述第一音频指纹单体和/或所述第二音频指纹单体进行加权。

10 前述的音频检索识别方法，其中，所述根据每个单独的所述第一音频指纹单体对多个已知音频进行第一排名包括：根据每个单独的所述第一音频指纹单体对多个已知音频进行词频-逆向文件频率 TF-IDF 排名。

前述的音频检索识别方法，其中，所述根据每个单独的所述第一音频指纹单体对多个已知音频进行词频-逆向文件频率 TF-IDF 方式的第一排名
15 包括：将所述已知音频的指纹索引与所述第一音频指纹单体进行匹配，以对所述已知音频进行所述 TF-IDF 排名。

前述的音频检索识别方法，其中，所述预先得到所述已知音频的指纹索引包括：根据所述第二强弱权重单体，预先得到所述已知音频的正排指纹索引和/或倒排指纹索引。

20 前述的音频检索识别方法，其中，所述将所述已知音频的指纹索引与所述第一音频指纹单体进行匹配包括：根据所述第一强弱权重单体，将所述音频的指纹索引与所述第一音频指纹单体进行绝对匹配。

前述的音频检索识别方法，其中，所述根据多个顺序排列的所述第一音频指纹单体对所述第一候选音频集合中的音频进行第二排名包括：根据
25 所述已知音频的指纹索引与所述第一音频指纹得到所述第一候选音频集合中的音频的相似度矩阵，根据所述相似度矩阵对所述第一候选音频集合中的音频进行排名。

前述的音频检索识别方法，其中，所述的根据所述已知音频的指纹索引与所述第一音频指纹得到所述第一候选音频集合中的音频的相似度矩阵，根据所述相似度矩阵对所述第一候选音频集合中的音频进行排名包括：
30 利用所述第一强弱权重单体和/或所述第二强弱权重单体进行加权，得到加权的所述相似度矩阵，并根据所述加权的相似度矩阵对所述第一候选音频集合中的音频进行排名。

前述的音频检索识别方法，其中，所述根据所述相似度矩阵对所述第一候选音频集合中的音频进行排名包括：根据所述相似度矩阵中的直线对
35 所述第一候选音频集合中的音频进行排名。

前述的音频检索识别方法，其中：所述获取待识别音频的音频指纹包括，获取所述待识别音频的多种类型的第一音频指纹；所述预先获取所述已知音频的音频指纹作为第二音频指纹包括，获取所述第一候选音频集合中的音频的多种类型的第二音频指纹；所述的根据所述已知音频的指纹索引与所述第一音频指纹得到所述第一候选音频集合中的音频的相似度矩阵包括，根据所述多种类型的第一音频指纹和所述多种类型的第二音频指纹确定所述相似度矩阵。

前述的音频检索识别方法，其中，每种类型的所述第一音频指纹包含多个第一音频指纹单体，每种类型的所述第二音频指纹包含多个第二音频指纹单体；所述的根据所述多种类型的第一音频指纹和所述多种类型的第二音频指纹确定所述相似度矩阵包括：分别确定同种类型的所述第二音频指纹单体与所述第一音频指纹单体之间的单体相似度，以得到多种所述单体相似度；根据所述多种单体相似度的平均值或最小值确定所述相似度矩阵。

前述的音频检索识别方法，还包括：预先对待识别音频和已知音频按照预设的时间长度切片，得到多段待识别子音频和多段已知子音频，对所述多段待识别子音频和所述多段已知子音频分别提取音频指纹，以得到长度相同的多个第一子音频指纹和多个第二子音频指纹。

前述的音频检索识别方法，还包括：在进行所述第一排名之前，对获得的待识别音频的所述第一音频指纹和已知音频的所述第二音频指纹按照预设的长度切片，以得到长度相同的多个第一子音频指纹和多个第二子音频指纹。

前述的音频检索识别方法，其中，所述多个第一音频指纹单体在所述第一音频指纹中按时间顺序排列，所述多个第二音频指纹单体在所述第二音频指纹中按时间顺序排列。

前述的音频检索识别方法，还包括：根据所述相似度矩阵确定所述待识别音频与所述识别结果中的音频的重复片段。

本公开的目的还采用以下技术方案来实现。依据本公开提出的音频检索识别装置，包括：音频指纹获取系统，用于获取待识别音频的音频指纹，其中，所述音频指纹包括用于表示所述待识别音频的内容特征的第一部分以及用于表示所述第一部分的可信程度的第二部分；检索识别系统，用于根据所述音频指纹对所述待识别音频进行识别，得到识别结果。

本公开的目的还可以采用以下的技术措施来进一步实现。

前述的音频检索识别装置，其还包括执行前述任一音频检索识别方法步骤的模块。

本公开的目的还采用以下技术方案来实现。依据本公开提出的一种音

频检索识别硬件装置，包括：存储器，用于存储非暂时性计算机可读指令；以及处理器，用于运行所述计算机可读指令，使得所述处理器执行时实现前述任意一种音频检索识别方法。

本公开的目的还采用以下技术方案来实现。依据本公开提出的一种计算机可读存储介质，用于存储非暂时性计算机可读指令，当所述非暂时性计算机可读指令由计算机执行时，使得所述计算机执行前述任意一种音频检索识别方法。

本公开的目的还采用以下技术方案来实现。依据本公开提出的一种终端设备，包括前述任意一种音频检索识别装置。

上述说明仅是本公开技术方案的概述，为了能更清楚了解本公开的技术手段，而可依照说明书的内容予以实施，并且为了让本公开的上述和其他目的、特征和优点能够更明显易懂，以下特举较佳实施例，并配合附图，详细说明如下。

附图说明

图1是本公开一个实施例的音频检索识别方法的流程框图。

图2是本公开一个实施例提供的获取音频指纹的流程框图。

图3是本公开一个实施例提供的对音频进行检索识别的流程框图。

图4是本公开一个实施例提供的第一排名的流程框图。

图5是本公开一个实施例提供的第二排名的流程框图。

图6是本公开一个实施例提供的利用动态规划法确定序列相似度评分的流程框图。

图7是本公开一个实施例提供的利用匀速音频法确定序列相似度评分的流程框图。

图8是本公开一个实施例提供的基于多种类型第一音频指纹、第二音频指纹确定相似度矩阵的流程框图。

图9是本公开一个实施例的音频检索识别装置的结构框图。

图10是本公开一个实施例提供的音频指纹获取系统的结构框图。

图11是本公开一个实施例提供的检索识别系统的结构框图。

图12是本公开一个实施例提供的第一排名模块的结构框图。

图13是本公开一个实施例提供的第二排名模块的结构框图。

图14是本公开一个实施例的基于多种类型第一音频指纹和第二音频指纹确定相似度矩阵的音频检索识别装置的结构框图。

图15是本公开一个实施例的音频检索识别硬件装置的硬件框图。

图16是本公开一个实施例的计算机可读存储介质的示意图。

图17是本公开一个实施例的终端设备的结构框图。

具体实施方式

为更进一步阐述本公开为达成预定发明目的所采取的技术手段及功效，以下结合附图及较佳实施例，对依据本公开提出的音频检索识别方法及装置的具体实施方式、结构、特征及其功效，详细说明如后。

图 1 为本公开的音频检索识别方法一个实施例的示意性流程图。请参阅图 1，本公开示例的音频检索识别方法，主要包括以下步骤：

步骤 S10，获取待识别音频（Query 音频）的音频指纹。其中，该音频指纹包括用于表示待识别音频的内容特征的第一部分以及用于表示该第一部分的可靠程度的第二部分。此后，处理进到步骤 S20。

步骤 S20，根据待识别音频的该音频指纹对待识别音频进行识别，得到识别结果。

本公开示例的音频检索识别方法，通过获取并利用音频对象的包括用于表示音频内容特征的第一部分和用于表示第一部分的可靠程度的第二部分的音频指纹特征来进行音频检索识别，能够提高音频检索识别的准确性、鲁棒性和效率。

下面对上述各步骤分别进行详细的陈述和说明。

一、关于步骤 S10。

图 2 为本公开一个实施例提供的获取音频指纹的示意性流程框图。由于对任何音频均可以按照图 2 所示的方法获取音频指纹，在本实施例的说明中不区分是否为待识别的音频。请参阅图 2，在本公开一个实施例中，前述的步骤 S10 的获取音频指纹的过程具体包括以下步骤：

步骤 S11，将音频转换成声谱图（Spectrogram）。具体地，通过短时傅里叶变换（Fast Fourier Transformation）将音频信号转换成时间-频率声谱图。其中的声谱图是一种常用的音频信号的二维频谱图，横轴是时间 t ，纵轴是频率 f ，图中每个点 (t, f) 的具体的取值 $E(t, f)$ 代表了信号的能量。需注意，对音频信号的具体类型不做限制，可以是静态文件（static file）也可以是流音频（streaming audio）。此后，处理进到步骤 S12。

在本公开的实施例中，可利用梅尔（MEL）变换对声谱图进行预处理，利用梅尔变换能够将频谱分成多个频率区块（频率 bin），而所分成的频率区块的数目是可以配置的。另外，还可以对声谱图进行人类听觉系统滤波（Human Auditory System filtering），利用人类听觉系统滤波等非线性变换，能够使得声谱图中的频谱分布更适合人耳感知。

需要说明的是，可以通过调整短时傅里叶变换中的各个超参数以适应不同的实际情况。在本公开的实施例中，可将步骤 S11 中的各个超参数设置为：在短时傅里叶变换中，时间窗设置为 100ms，间隔设置为 50ms；在

梅尔变换中，频率区块的数目设置为 32~128。

步骤 S12，确定声谱图中的特征点。

具体地，采用多种标准中的一种来确定特征点，例如，可以将特征点选为声谱图中的能量的极大值点，或者，也可以选为能量的极小值点。其中，如果声谱图中的一个点 (t, f) 的能量 $E(t, f)$ 能够同时满足：
5 $E(t, f) > E(t+1, f)$ 、 $E(t, f) > E(t-1, f)$ 、 $E(t, f) > E(t, f+1)$ 且 $E(t, f) > E(t, f-1)$ ，则该 (t, f) 点为声谱图中的能量极大值点。类似地，如果一个点 (t, f) 的能量 $E(t, f)$ 能够同时满足： $E(t, f) < E(t+1, f)$ 、 $E(t, f) < E(t-1, f)$ 、 $E(t, f) < E(t, f+1)$ 且 $E(t, f) < E(t, f-1)$ ，则该 (t, f) 点
10 为声谱图中的能量极小值点。此后，处理进到步骤 S12。

在本公开的实施例 中，由于选取能量极值点作为特征点存在：能量极值点易受噪声影响；不易控制极值点的数量，可能一个声谱图中没有极值点，而另一个声谱图中有多个极值点，导致特征点不均匀；需要存储额外的时间戳以记录能量极值点在声谱图中的位置等问题。因此，也可以不选
15 能量的极值点作为特征点，而是选取固定点作为特征点，例如可以选取频率值与预设的频率设定值相等的点（频率固定的点）。进一步地，可按照频率大小预设低频、中频、高频的多个频率设定值（低频、中频、高频的具体值是可以设置的）。通过选取频率为低频、中频、高频的多个固定点作为特征点，可以使得选取的特征点更加均匀。需要注意的是，也可以按照其
20 他标准选取固定点，如选取与一个或多个预设能量值相等的点。

需要说明的是，可以通过调整所选取的特征点的数量以适应不同的实际情况。在本公开的实施例 中，可将步骤 S12 中的超参数设置为：特征点的密度设置为每秒 20~80 个。

步骤 S13，在声谱图上，在特征点的附近，为特征点确定一个或多个掩模（mask），每个掩模包含（或者说，覆盖）多块声谱图上的区域（不妨称为谱区域）。此后，处理进到步骤 S14。
25

具体地，在声谱图中，每个掩模所包含的多块谱区域可以是对称分布的：

以时间轴对称（即，多个谱区域具有相同的频率范围），例如，在一个梅尔-声谱图中，可以为特征点确定一个包含 R11 和 R12 这两块谱区域的掩模，R11、R12 均位于特征点的左侧，且 R11 位于 R12 的左侧，并且 R11 与 R12 覆盖相同的频率区块；
30

或者以频率轴对称（即，多个谱区域具有相同的时间范围）。例如，在一个梅尔-声谱图中，可以为特征点确定一个包含 R13 和 R14 这两块谱区域的掩模，R13 位于特征点的上侧，R14 位于特征点的下侧，并且 R13 与 R14 具有相同的时间范围；
35

或者以特征点为中心而中心对称分布，例如，在一个梅尔-声谱图中，可以为特征点确定一个包含 R15 和 R16 这两块谱区域的掩模，R15 位于特征点的左上侧，R16 位于特征点的右下侧，并且 R15 与 R16 以特征点为中心而相互对称。

5 当然，一个掩模所包含的多块谱区域也可以同时满足多种对称分布情况。例如，可以为特征点确定一个包含 R21、R22、R23 和 R24 这四块谱区域的掩模，R21、R22、R23、R24 分别位于特征点的左上、右上、左下、右下，并且 R21 与 R22 具有相同的频率范围、R23 与 R24 具有相同的频率范围、R21 与 R23 具有相同的时间范围、R22 与 R24 具有相同的时间范围，而且这
10 四块谱区域还以特征点为中心而中心对称。需要说明的是，一个掩模的四个谱区域并非一定以特征点为中心而中心对称分布，例如，可以均位于特征点的左侧，且在频率轴上分布于特征点的两侧。

需要说明的是，属于同一掩模的多块谱区域之间是可以相互交叠的。另外，不同的掩模之间也是可以相互交叠的。可选地，每个掩模可包含偶
15 数个谱区域。

需要注意的是，掩模可以是按照固定的预设标准确定的，即每个掩模在声谱图中的位置及覆盖的区域是预先设置好的。或者，也可以不预先固定掩模的位置和范围，而是使用数据驱动的方式自动确定掩模区域：从大量掩模中选取协方差最小、最有区分度的掩模。

20 步骤 S14，确定每个谱区域的均值能量。具体地，对于仅包含一个点的谱区域，该谱区域的均值能量就是这个点的能量值；当谱区域由多个点组成时，可以将该谱区域的均值能量设置为这多个点的能量值的平均值。此后，处理进到步骤 S15。

步骤 S15，根据掩模中的多块谱区域的均值能量，确定音频指纹比特
25 (bit)。需要注意的是，该音频指纹比特即为前述的音频指纹中的用于表示音频的内容特征的第一部分。此后，处理进到步骤 S16。

在本公开实施例的步骤 S15 中，可根据一个掩模所包含的多个谱区域的均值能量的差值确定一个音频指纹比特。

具体地，如果一个掩模包含两个谱区域，例如前述的包含 R11 和 R12
30 两块谱区域的示例，可以按照下面的公式一来计算 R11、R12 的均值能量的差值 D1：

$$D1 = E(R11) - E(R12), \quad (\text{公式一})$$

然后判断差值 D1 的正负，如果差值 D1 为正值，则得到一个取值为 1 的音频指纹比特，如果差值 D1 为负值，则得到一个取值为 0 的音频指纹比特。

35 如果一个掩模包含四个谱区域，例如前述的包含 R21、R22、R23、R24 四块谱区域的示例，可以按照下面的公式二来计算 R21、R22、R23、R24 的

均值能量的差值 D2:

$$D2 = (E(R21) + E(R22)) - (E(R23) + E(R24)), \quad (\text{公式二})$$

然后判断差值 D2 的正负, 如果差值 D2 为正值, 则得到一个取值为 1 的音频指纹比特, 如果差值 D2 为负值, 则得到一个取值为 0 的音频指纹比特。

5 需要说明的是, 并非必须通过差值 D2 来确定一个包含四个谱区域的掩模的音频指纹比特, 也可以利用其他形式的差值来确定音频指纹比特。例如, 也可以计算这四个谱区域的均值能量的二阶差值 D3:

$$D3 = (E(R23) - E(R24)) - (E(R21) - E(R22)), \quad (\text{公式三})$$

然后判断差值 D1 的正负来确定音频指纹比特。

10 需要说明的是, 如果为特征点确定了多个掩模, 则能够对应地得到多个音频指纹比特。

步骤 S16, 确定音频指纹比特对应的强弱权重比特, 该强弱权重比特用于表示该音频指纹比特的可信程度。需要注意的是, 该强弱权重比特即为前述的音频指纹中的用于表示第一部分的可信程度的第二部分。具体他, 15 将可信度高的音频指纹比特定义为强比特, 将可信度低的音频指纹比特定义为弱比特。判断一个音频指纹比特的可信程度, 并根据该音频指纹比特是强比特还是弱比特来确定强弱权重比特的取值。此后, 处理进到步骤 S17。

在本公开的实施例中, 如果音频指纹比特是根据一个掩模所包含的多个谱区域均值能量的差值来确定的, 则步骤 S16 具体包括: 判断生成该音频指纹比特所使用的该差值的绝对值是否达到 (或超过) 预设的强弱比特 20 阈值; 如果达到强弱比特阈值, 则将该音频指纹比特确定为强比特, 并得到一个与该音频指纹比特对应的取值为 1 的强弱权重比特; 如果未达到强弱比特阈值, 则将该音频指纹比特确定为弱比特, 并得到一个与该音频指纹比特对应的取值为 0 的强弱权重比特。

25 作为一个具体示例, 如果一个音频指纹比特是通过判断前述公式二的四个谱区域均值能量的差值 D2 的正负来确定的, 则步骤 S16 包括: 判断该差值 D2 的绝对值与预设的强弱比特阈值 T 的大小关系, 如果 $|D2| \geq T$, 则该音频指纹比特是强比特, 并将该音频指纹比特对应的强弱权重比特取值设置为 1; 如果 $|D2| < T$, 则该音频指纹比特是弱比特, 并将该音频指纹比特对应的强弱权重比特取值设置为 0。需要说明的是, 该强弱比特阈值可以是多种类型的阈值: 该强弱比特阈值可以是预设的固定值, 例如可以固定取为 1; 或者, 该强弱比特阈值也可以是基于均值能量的差值而得到的数值, 例如可将该强弱比特阈值设置为多个掩模 (或多个特征点) 对应的多个差值的平均数 (事实上不限于平均数, 也可以是任意一个介于最大的差值与最小的差值之间的数值), 并且将差值达到该平均数的音频指纹比特确定 30 为强比特, 将差值未达到该平均数的音频指纹比特确定为弱比特; 再或

者，该强弱比特阈值也可以是个比例值，例如可将该强弱比特阈值设置为60%，在多个掩模（或多个特征点）对应的多个差值中，如果一个差值的绝对值位于所有差值中的前60%，则将该音频指纹比特确定为强比特，否则将该音频指纹比特确定为弱比特。

5 步骤 S17，根据该音频指纹比特和该强弱权重比特确定该音频的音频指纹。具体地，对音频指纹的组合方式、音频指纹的长度不做限制，仅需使得音频指纹包括一个或多个特征点所对应的各个音频指纹比特（以形成音频指纹的第一部分）和各个强弱权重比特（以形成音频指纹的第二部分）。在本公开的一些实施例中，该音频指纹包含多个音频指纹单体以及与各个
10 该音频指纹单体相对应的强弱权重单体，该音频指纹单体包含音频的多个该音频指纹比特，该强弱权重单体包含与该多个音频指纹比特相对应的多个该强弱权重比特；例如，可以将一个特征点的所有掩模所对应的音频指纹比特组合在一起而得到一个音频指纹比特序列作为音频指纹单体，将对应的强弱权重比特组合在一起而得到与该音频指纹比特序列长度相等的强
15 弱权重比特序列作为强弱权重单体，将多个特征点所对应的音频指纹单体、强弱权重单体按特征点的时间顺序排列以组成音频指纹。可选地，获得的音频指纹单体的长度可以是 32bits。

本公开通过在提取音频指纹比特的同时，提取该音频指纹比特对应的强弱权重比特，能够为一段音频生成一个准确性高、鲁棒性好的音频指纹。

20 可选地，本公开的步骤 S10 还包括：为音频指纹添加一个时间戳字段，用于表示音频起始位置与该特征点的时间差的字段，该字段可以是一个 hash 值。而如果将特征点设为固定点，则可以不包含本步骤，即不必记录该时间戳。

可选地，本公开的步骤 S10 还包括：为音频指纹添加一个音频标识字段，用于记录该音频指纹所对应的音频的 ID 标识信息，该字段可以是一个
25 hash 值。

可选地，本公开的步骤 S10 还包括：将原始音频按时间分成多段子音频；按照前述方法的步骤，对各段子音频提取音频指纹，得到多个音频指纹；将提取的各段子音频的音频指纹组合在一起，得到该整段音频的音频
30 指纹。

二、关于步骤 S20。

为了便于叙述和理解，不妨将待识别音频的音频指纹称为第一音频指纹，第一音频指纹所包含的音频指纹单体称为第一音频指纹单体，第一音频指纹单体对应的强弱权重单体称为第一强弱权重单体。

35 图 3 为本公开一个实施例提供的根据音频指纹对音频进行检索识别的示意性流程框图。请参阅图 3，在本公开一个实施例中，前述的步骤 S20 的

对待识别音频进行检索识别的过程具体包括以下步骤：

步骤 S21, 根据该第一音频指纹, 对多个已知音频进行第一排名, 根据该第一排名的结果, 取出前 k 个已知音频作为第一候选音频集合。其中的 k 为正整数, 而 k 的具体取值是可以设置的。具体地, 该第一排名是根据每个单独的第一音频指纹单体与已知音频的匹配情况进行的排名。进一步地, 该第一排名可以是根据各个第一音频指纹单体对已知音频进行的词频-逆向文件频率排名 (term frequency - inverse document frequency ranking, 简称为 TF-IDF 排名)。此后, 处理进到步骤 S22。

步骤 S22, 根据该第一音频指纹, 对该第一候选音频集合进行第二排名, 根据该第二排名的结果, 取出第一候选音频集合中的前 n 个第一候选音频作为识别结果。其中的 n 为正整数, 而 n 的具体取值是可以设置的。具体地, 该第二排名为根据多个顺序排列的第一音频指纹单体, 对该第一候选音频集合中的音频进行的排名。例如, 该多个顺序排列的第一音频指纹单体包括第一音频指纹中的连续的一部分、该第一音频指纹整体, 和/或该多个顺序排列的第一音频指纹单体包括第一音频指纹中的具有相同间隔的序号的多个第一音频指纹单体, 例如序号为 1、3、5、7、... 的多个第一音频指纹单体。

从而根据该识别结果在 Meta 数据库中进行检索, 能够得到该识别结果的音频信息, 例如识别出的音频的名称、作者、出处等等。当识别结果包括多个音频时, 可以同时提供多个识别出的音频的信息。

在本实施例中, 在进行步骤 S21 的第一排名和/或步骤 S22 的第二排名的过程中, 在利用到音频指纹时, 可以根据音频指纹中的强弱权重单体的对音频指纹单体进行加权。由于不加权的第一排名、第二排名过程相当于对在排名时对每个音频指纹单体施加了相同的权重, 因此以下仅对利用强弱权重对音频指纹进行加权的第一排名和第二排名的过程进行具体说明。

本公开提出的媒体检索方法, 通过进行第一排名和第二排名以得到检索结果, 能够大大提高媒体检索的准确性和效率。

关于前述的步骤 S21。

前述的已知音频可以是一个音频数据库中的音频。在该音频数据库中存储有已知音频的音频指纹, 并且在存储的已知音频的音频指纹中包含有与第一音频指纹利用相同提取方法得到的与第一音频指纹相同类型的音频指纹, 从而已知音频的音频指纹中也包括用于表示音频的内容特征的第一部分以及用于表示该第一部分的可信程度的第二部分。

在本公开的一些实施例中, 本公开的音频检索识别方法还包括: 预先获取多个已知音频的音频指纹, 为了便于叙述和理解, 不妨将已知音频的音频指纹称为第二音频指纹, 第二音频指纹所包含的音频指纹单体称为第

二音频指纹单体，第二音频指纹所包含的强弱权重单体称为第二强弱权重单体；对该第二音频指纹进行索引，以预先得到已知音频的指纹索引；将该指纹索引与待识别音频的第一音频指纹单体进行匹配，以对多个已知音频进行 TF-IDF 排名。

5 具体地，前述的预先得到已知音频的指纹索引进一步包括，预先得到已知音频的音频指纹的正排指纹索引（forward index）和倒排指纹索引（inverted index），以便于对音频指纹的检索和比对。该正排指纹索引和倒排指纹索引可以预先存储在音频数据库中。其中，正排指纹索引用于记录各个已知音频的音频指纹，即记录了各个已知音频的音频指纹具体包含
10 了哪些音频指纹单体以及这些音频指纹单体的顺序；倒排指纹索引用于记录各个音频指纹单体在哪个或哪些已知音频的音频指纹中出现。具体地，可以利用键值对（key-value 对）的形式来存储该正排指纹索引和倒排指纹索引：在正排指纹索引中，用一个键（key）表示一个音频的编号（或者，称为音频 ID），而与该键对应的值（value）记录该音频包含了哪些音频指
15 纹单体以及这些音频指纹单体的顺序，不妨将正排指纹索引中的键、值分别称为正排键、正排值；在倒排指纹索引中，用一个键（key）表示一个音频指纹单体，而与该键对应的值（value）记录包含有该音频指纹单体的音频的编号，不妨将倒排指纹索引中的键、值分别称为倒排键、倒排值。

值得注意的是，可以根据强弱权重对第二音频指纹进行索引，以提高
20 鲁棒性。具体地，在确定正排指纹索引的过程中，可以在正排指纹索引中记录已知音频的各个音频指纹单体对应的强弱权重单体。在确定倒排指纹索引的过程中，在确定一个待索引音频指纹单体是否出现在一个已知音频中时，可以忽略该待索引音频指纹单体中的弱比特，而仅判断该待索引音频指纹单体中的所有强比特是否与该已知音频中的某个音频指纹单体的对
25 应比特相一致；例如，如果一个待索引音频指纹单体中的第一个、第三个音频指纹比特为强比特、其他的音频指纹比特为弱比特，则该待索引音频指纹单体的倒排指纹索引记录有：包含与待索引音频指纹单体具有相同的第一、第三音频指纹比特的音频指纹单体的已知音频编号。

其中的 TF-IDF 排名是一类通过对信息进行词频和逆向文件频率加权，
30 来判断信息的重要程度，以进行排名的技术。其中的词频是指一个词（或者说，一个信息）在某个文章（或者说，某个文件）中出现的频率，词频越高说明该词对于该文章越重要；其中的文件频率是指一个词出现在了文章库中的多少个文章中，而逆向文件频率是文件频率的倒数（实际计算时，还可对逆向文件频率取对数，或者定义逆向文件频率是文件频率的倒数的
35 对数），逆向文件频率越高，说明该词的区分度越好。因此，TF-IDF 排名利用词频与逆向文件频率的乘积的大小进行排名。事实上，可以将一个音频

的音频指纹作为一个文章，而每个音频指纹单体作为一个词，从而能够利用 TF-IDF 方式对已知音频进行排名。

另外，如果对音频数据库中的所有已知音频都进行第一排名，可能会影响检索识别的效率，因此在第一排名之前，可以先对音频数据库中的已知音频进行绝对匹配 (exact match)。其中的绝对匹配，用于选出所包含的第一音频指纹单体的数量在预设数量或预设比例以上的已知音频作为第二候选音频集合。然后再对该第二候选音频集合进行第一排名，以选出第一候选音频集合。

图 4 为本公开一个实施例提供的第一排名的示意性流程框图。请参阅图 4，在本公开一个实施例中，第一排名具体包括以下步骤：

步骤 S31，根据倒排指纹索引，统计各个第一音频指纹单体在哪些已知音频的第二音频指纹中出现，以从音频数据库中匹配出包含预设数量以上第一音频指纹单体的已知音频作为第二候选音频集合。值得注意的是，在匹配过程中，可以根据一个第一音频指纹单体对应的强弱权重单体，仅判断该第一音频指纹单体中的强比特在已知音频的第二音频指纹中的匹配情况，而忽略该第一音频指纹单体中的弱比特的匹配情况，以提高鲁棒性。此后，处理进到步骤 S32。

需要注意的是，“预设数量以上第一音频指纹单体”中的“数量”指的是第一音频指纹单体的种类。具体地，该预设数量可以是一个，从而匹配出的第二候选音频集合为第二音频指纹中至少出现了某一种第一音频指纹单体的已知音频；该预设数量也可以是多个，不妨为 p 个 (p 为正整数)，从而匹配出的第二候选音频集合为第二音频指纹中至少出现了 p 种第一音频指纹单体的已知音频。

步骤 S32，基于正排指纹索引，确定一个第一音频指纹单体在一个第二候选音频的第二音频指纹中的词频。该词频为：一个第一音频指纹单体在一个第二音频指纹所包含的全部音频指纹单体之中所占的比例。值得注意的是，该正排指纹索引可以是前述的根据强弱权重得到的指纹索引。此后，处理进到步骤 S33。

步骤 S33，基于倒排指纹索引，确定一个第一音频指纹单体的文件频率。该文件频率为：在多个已知音频之中（例如，可以是音频数据库中所有的已知音频），第二音频指纹中包含有该第一音频指纹单体的已知音频的数量占已知音频总数的比例。值得注意的是，该倒排指纹索引可以是前述的根据强弱权重得到的指纹索引。此后，处理进到步骤 S34。

步骤 S34，根据各个第一音频指纹单体在一个第二候选音频的第二音频指纹中的词频以及各个第一音频指纹单体的文件频率，确定该第二候选音频的词频-逆向文件频率评分。此后，处理进到步骤 S35。

步骤 S35, 根据得到的各个第二候选音频的词频-逆向文件频率评分对第二候选音频集合进行排名, 得到第一排名的结果, 从该第一排名结果中取出前 k 个第二候选音频作为第一候选音频集合。同时, 还可以返回各个第一候选音频的第二音频指纹 (正排指纹索引), 以备在后续的第二排名中
5 基于该第二音频指纹对第一候选音频集合进行进一步处理。

在本实施例中, 可以利用索引服务器, 将待识别音频的第一音频指纹单体的集合作为索引请求, 根据前述的正排指纹索引和倒排指纹索引, 进行绝对匹配和 TF-IDF 排名, 以召回第一候选音频集合并同时返回得到的各个第一候选音频的正排指纹索引。具体地, 可以利用开源的 Elasticsearch
10 搜索引擎进行上述的各个步骤, 以达到快速检索的效果。

值得注意的是, 绝对匹配和第一排名着重关注各个第一音频指纹单体出现在哪些已知音频中以及第一音频指纹单体本身的检索情况, 并未考虑各个第一音频指纹单体在第一音频指纹中的顺序对检索的影响, 或者说并未考虑音频指纹的整体或连续多个音频指纹单体的检索情况。

15 本公开提出的音频检索识别方法, 根据包含强弱权重的音频指纹来进行绝对匹配和基于 TF-IDF 方式的第一排名, 能够大大提高音频检索识别的准确性和效率。

关于前述的步骤 S22。

在本公开的一些实施例中, 该第二排名为根据多个顺序排列的第一音频指纹单体所组成的具有先后顺序的序列在第一候选音频的音频指纹中出现的情况, 对该第一候选音频集合中的音频进行的排名。具体地, 该第二排名包括: 根据已知音频的指纹索引与第一音频指纹得到该第一候选音频集合中的音频的相似度矩阵, 根据该相似度矩阵对该第一候选音频集合中的音频进行排名。值得注意的是, 在确定相似度矩阵的过程中, 可以根据
20 第一音频指纹对应的强弱权重和/或已知音频的指纹索引中的强弱权重进行加权, 并利用加权后的相似度矩阵对第一候选音频集合中的音频进行排名, 以提高鲁棒性。

图 5 为本公开一个实施例提供的第二排名的示意性流程框图。请参阅图 5, 在本公开一个实施例中, 该第二排名具体包括以下步骤:

30 步骤 S41, 获取第一候选音频集合中的一个第一候选音频 (事实上每个第一候选音频都是已知音频) 的第二音频指纹。具体地, 可以根据已知音频的指纹索引 (例如, 正排指纹索引) 获取该第二音频指纹。不妨假设待识别音频的第一音频指纹包含 M_1 个第一音频指纹单体, 该第一候选音频的第二音频指纹包含 M_2 个第二音频指纹单体, 其中的 M_1 和 M_2 为正整数。
35 在本公开的一些示例中, 第一音频指纹中包含有与各个第一音频指纹单体对应的强弱权重单体 (不妨称之为第一强弱权重单体), 和/或第二音频指

纹中包含有与各个第二音频指纹单体对应的强弱权重单体（不妨称之为第二强弱权重单体）。此后，处理进到步骤 S42。

步骤 S42, 确定该第一候选音频的第二音频指纹所包含的各个第二音频指纹单体与各个第一音频指纹单体之间的单体相似度，得到 $M_1 \times M_2$ 个单体相似度。每个单体相似度表示一个第一音频指纹单体与一个第二音频指纹单体之间的相似程度，具体可以是，单体相似度越大表示越相似。值得注意的是，在确定该单体相似度的过程中，可以根据第一强弱权重单体和/或第二强弱权重单体，对各个第一音频指纹单体、第二音频指纹单体进行加权，然后根据加权后的第一、第二音频指纹单体确定该单体相似度。在本公开的一种示例中，由于音频数据库中的数据信息的准确性更高，可以利用第二强弱权重单体分别对第一音频指纹单体、第二音频指纹单体进行加权。此后，处理进到步骤 S43。

在本公开的实施例中，可以根据音频指纹的类型，选择能够判断两个音频指纹单体的相似程度的距离或度量作为该单体相似度。具体地，当第一音频指纹单体、第二音频指纹单体同为按照前述实施例中的步骤 S11 至步骤 S17 方法得到的二值指纹时，计算第一音频指纹单体、第二音频指纹单体之间的汉明距离（Hamming 距离），再计算音频指纹单体的长度（比特数）与该汉明距离的差值，并将该差值与该音频指纹单体长度的比值确定为单体相似度，用以表示两个二值指纹中的相同比特所占的比例。其中的汉明距离是一种信息论领域中常用的度量，两个等长字符串之间的汉明距离是两个字符串对应位置的不同字符的个数。在实际计算汉明距离时，可以对两个字符串进行异或运算，并统计结果为 1 的个数，而这个数就是汉明距离。需要说明的是，利用同种方法提取得到的音频指纹单体具有相同的长度。而本公开的利用强弱权重对这种汉明距离类型的单体相似度进行加权的具體方法为，先利用强弱权重单体中的强弱权重比特对音频指纹单体中的对应的音频指纹比特进行加权，再对第一音频指纹单体、第二音频指纹单体进行异或运算，以得到利用强弱权重加权的单体相似度。另外需要说明的是，不限于利用汉明距离表示该单体相似度，而是可以利用任何可以判断两个音频指纹单体的相似程度的距离或度量。

步骤 S43, 根据各个单体相似度，确定该第一候选音频与待识别音频之间的相似度矩阵（similarity matrix）。

具体地，该相似度矩阵中的每个点对应一个单体相似度，使得该相似度矩阵记录有一个第一候选音频的各个第二音频指纹单体与各个第一音频指纹单体之间的单体相似度。并且，该相似度矩阵的各个点：在横向上按照待识别音频的各个第一音频指纹单体在第一音频指纹中的先后顺序排列，且在纵向上按照第一候选音频的各个第二音频指纹单体在第二音频指

纹中的先后顺序排列。从而位于第 i 行第 j 列的点表示待识别音频的第 i 个第一音频指纹单体和第一候选音频的第 j 个第二音频指纹单体之间的单体相似度，进而该相似度矩阵为一个 $M_1 \times M_2$ 矩阵。此后，处理进到步骤 S44。

5 需要说明的是，在实际操作中，并非一定先进行步骤 S42 的计算各个单体相似度，再进行步骤 S43 的确定相似度矩阵，而是可以直接确定相似度矩阵，在确定该相似度矩阵的各个点的过程中计算对应的单体相似度。

步骤 S44，根据每个第一候选音频的相似度矩阵，确定该第一候选音频的序列相似度评分。该序列相似度评分用于表现该第一候选音频与待识别
10 音频之间的相似程度。该序列相似度评分可以是一个 0 到 1 之间的分数，数字越大表示两段音频越相似。此后，处理进到步骤 S45。

具体地，根据相似度矩阵中的直线来确定该的序列相似度评分。

需注意，由于音频指纹一般包含有穷的多个音频指纹单体，从而相似度矩阵为有穷矩阵，因此实际上所谓的“直线”是相似度矩阵中的多个点
15 组成的有穷长的线段。该直线具有斜率，该斜率为直线所包括的多个点的连线的斜率。另外，该直线的起点和终点可以是相似度矩阵中的任意的点，不必是位于边缘的点。

本公开所说的直线包括相似度矩阵中的对角线、与该对角线相平行的各条线段这些在相似度矩阵中从左下到右上的斜率为 1 的直线，还包括斜
20 率不为 1 的直线。例如，可以是斜率近似于 1 的直线，以提高音频检索识别的鲁棒性；可以是斜率为 2、3、... 或 $1/2$ 、 $1/3$ 、... 等的直线，以应对经过调速的音频的检索识别；甚至可以是斜率为负数的直线（在相似度矩阵中从左上到右下的直线），以应对经过反向播放处理的音频的检索识别。其中的对角线为由位于 $(1,1)$ 、 $(2,2)$ 、 $(3,3)$... 的点组成的线段（事
25 实上就是以左上角的点为起点且斜率为 1 的一条直线）。

事实上，相似度矩阵中的每条直线均由顺序排列的多个单体相似度构成，因此由于每条直线表现了多个顺序排列的音频指纹单体对的相似情况，从而能够表现待识别音频中的一段音频片段与已知音频中的一段音频片段的相似程度。其中每个音频指纹单体对包括一个第一音频指纹单体和一个
30 第二音频指纹单体（也就是说，每条直线表现了多个顺序排列的第一音频指纹单体与多个顺序排列的第二音频指纹单体之间的相似程度）。而直线的斜率、起点终点表现了两段音频片段的长度、位置。例如，由 $(1,1)$ 、 $(2,3)$ 、 $(3,5)$ 、 $(4,7)$ 构成的直线，由于表现了序数为 1 的第一音频指纹单体与序数为 1 第二音频指纹单体之间的相似情况、序数为 2 的第一音频指纹单体与序数为 3 第二音频指纹单体之间的相似情况、...，从而该直线能够反应
35 序数为 1、2、3、4 的第一音频指纹单体所对应的一段待识别音频片段与序

数为 1、3、5、7 的第二音频指纹单体所对应的一段已知音频片段之间的相似情况。

因此，可以根据相似度矩阵中的直线来确定一个第一候选音频与待识别音频之间的相似情况：不妨将一个直线所包含的各个单体相似度的平均情况（或总体情况）定义为该直线的直线相似度，该直线相似度能够体现
5 对应的多个第一音频指纹单体与多个第二音频指纹单体之间的相似情况；在相似度矩阵中确定一条直线相似度最高的直线，不妨称为匹配直线；将匹配直线的直线相似度确定为第一候选音频的序列相似度评分。

需要注意的是，在确定匹配直线的过程中，可以从预设的多条直线
10 中确定一条直线相似度最高的直线，例如该预设的多条直线为所有的斜率为预设的斜率设定值（比如斜率为 1）的直线，或者，也可以是从相似度矩阵中选取使得单体相似度的大小排名靠前的多个点，再根据这些点拟合成一条直线，以生成一条使得直线相似度相对最高的直线。

步骤 S45，根据各个第一候选音频的该序列相似度评分对第一候选音频
15 集合进行排名，得到第二排名的结果，从该第二排名结果中取出前 n 个第一候选音频作为识别结果。

本公开提出的音频检索识别方法，根据包含强弱权重的音频指纹并基于相似度矩阵来进行第二排名，能够大大提高音频检索识别的准确性和效率。

在本公开的一个具体实施例中，可以利用动态规划法来根据相似度矩阵确定序列相似度评分。图 6 为本公开一个实施例提供的利用动态规划法
20 进行音频检索识别的示意性流程框图。请参阅图 6，在一种实施例中，步骤 S44 包括以下具体步骤：

步骤 S44-1a，将相似度矩阵中的斜率为预设的斜率设定值的多条直线
25 定义为备选直线，根据每条备选直线所包含的各个单体相似度确定该备选直线的直线相似度。具体地，一条直线的直线相似度可以设置为该直线所包含的各个单体相似度的平均值，或者可以设置为该直线所包含的各个单体相似度的总和值。在一种具体示例中，可以将斜率设定值取为 1，即前述的备选直线为：相似度矩阵中的对角线以及与该对角线平行的直线。此后，
30 处理进到步骤 S44-1b。

需要注意的是，在本公开的一种实施例中，步骤 S44-1a 还包括：先从
备选直线中排除那些包含的单体相似度的数量少于预设的直线长度设定值的直线，然后再进到步骤 S44-1b。或者说，在本实施例中，备选直线还须
35 满足：包含的单体相似度的数量达到预设的直线长度设定值。通过排除单体相似度过少的直线，可以排除当直线包含的单体相似度过少而影响最终得到的序列相似度评分的准确性的问题。

步骤 S44-1b, 从该多条备选直线中, 确定一条使得该直线相似度最大的备选直线, 并定义为第一匹配直线。此后, 处理进到步骤 S44-1c。

步骤 S44-1c, 将该第一匹配直线的直线相似度确定为序列相似度评分。

需要注意的是, 在本公开的一些实施例中, 步骤 S44-1a 中的预设的斜率设定值可以为多个, 即备选直线为斜率与多个斜率设定值中任意一个相等的直线, 例如备选直线可以为斜率为 1、-1、2、1/2 等的直线, 并且在步骤 S44-1b 中, 从斜率为多个斜率设定值中任意一个的多条备选直线中确定一条第一匹配直线。

本公开提出的音频检索识别方法, 通过利用动态规划法来确定序列相似度评分, 能够提高音频检索识别的准确性和效率。

在本公开的另一个具体实施例中, 可以利用匀速媒体法来根据相似度矩阵确定序列相似度评分。图 7 为本公开一个实施例提供的利用匀速媒体法进行音频检索识别的示意性流程框图。请参阅图 7, 在一种实施例中, 步骤 S34 包括以下具体步骤:

步骤 S44-2a, 在相似度矩阵中选取单体相似度最大的多个点作为相似度极值点。所取的相似度极值点的具体数量可以是预设的。此后, 处理进到步骤 S44-2b。

步骤 S44-2b, 基于该多个相似度极值点, 在该相似度矩阵中拟合出一条直线作为第二匹配直线。在一些具体示例中, 基于该多个相似度极值点拟合出一条具有预设的斜率设定值或接近预设的斜率设定值的直线作为第二匹配直线, 例如, 拟合出一条斜率接近 1 的直线。具体地, 可以利用随机抽样一致法 (Random Sample Consensus 法, 简称为 RANSAC 法) 在该相似度矩阵中拟合出一条斜率接近斜率设定值的直线。其中的 RANSAC 法是一种常用的根据一组包含异常数据的样本数据集, 计算出数据的数学模型参数, 以得到有效样本数据的方法。此后, 处理进到步骤 S44-2c。

步骤 S44-2c, 根据该第二匹配直线所包含的多个单体相似度来确定序列相似度评分。具体地, 可将该第二匹配直线上的各个单体相似度的平均值确定为该序列相似度评分。

本公开提出的音频检索识别方法, 通过利用匀速媒体法来确定序列相似度评分, 能够提高音频检索识别的准确性和效率。

进一步地, 其中相似度矩阵可以是由多种音频相似度综合考量得到的。具体地, 本公开的音频检索识别还包括: 获取待识别音频的多种类型的第一音频指纹, 获取第一候选音频集合中的音频的多种类型的第二音频指纹, 根据基于多种类型的第二音频指纹得到的指纹索引以及多种类型的第一音频指纹来确定相似度矩阵。

图 8 为本公开一个实施例的基于多种类型的第一音频指纹和第二音频

指纹确定相似度矩阵以进行音频检索的示意性流程框图。请参阅图 8，在本公开的一个实施例中，本公开的音频检索识别方法包括：

步骤 S51，利用多种音频指纹提取方法，获取待识别音频的多种类型的第一音频指纹，每种类型的第一音频指纹包含多个用于表示音频内容特征的第一部分不妨称为第一音频指纹单体。可选地，至少一些类型的第一音频指纹包含用于表示第一部分的可信程度的第二部分。例如，同时获取待识别音频的按照前述实施例中的步骤 S11 至步骤 S17 方法得到的音频指纹、以及其他类型的音频指纹。此后，处理进到步骤 S52。

步骤 S52，获取一个已知音频（具体地，可以是前述的第一候选音频集合中的音频）的多种类型的第二音频指纹，每种类型的第二音频指纹包含多个用于表示音频内容特征的第一部分不妨称为第二音频指纹单体。可选地，至少一些类型的第二音频指纹包含用于表示第一部分的可信程度的第二部分。例如，同时获取已知音频的按照前述实施例中的步骤 S11 至步骤 S17 方法得到的音频指纹、以及其他类型的音频指纹。此后，处理进到步骤 S53。

步骤 S53，利用与前述实施例的步骤 S42 相似的方法，分别确定同种类型的该第二音频指纹单体与该第一音频指纹单体之间的单体相似度。从而对应于多种类型的音频指纹，能够得到一个已知音频的多种单体相似度。此后，处理进到步骤 S54。

步骤 S54，确定多种单体相似度的平均值或最小值；根据多种单体相似度的该平均值或该最小值，利用前述实施例的步骤 S43 相似的方法确定该已知音频的相似度矩阵。

此后，处理进到前述示例的步骤 S44，并在步骤 S44 中根据该基于多种单体相似度的平均值或最小值的相似度矩阵，来确定序列相似度评分以及确定第二排名的结果等步骤。

利用多种相似度的平均值或最小值确定相似度矩阵的效果在于：利用单种音频指纹得到的相似度进行音频检索识别可能存在误匹配的情况，通过取多种音频指纹的相似度的平均值或取最小值，能够减少或排除该误匹配问题，从而提高音频检索识别的准确性。

需要说明的是，在取多种单体相似度的平均值或最小值之前，需要确保各种单体相似度具有一致的取值范围，例如可以预先将所有类型的单体相似度的取值范围均设置为 0 到 1。事实上，前述的根据汉明距离确定的单体相似度的示例已将单体相似度的取值范围设置为 0 到 1。

在本公开的一些实施例中，该音频检索识别方法还包括：在进行第一排名之前，对获取的待识别音频的第一音频指纹以及已知音频的第二音频指纹按照预设的固定长度切片，得到多个长度相同（包含相同数量的音频

指纹单体)的第一子音频指纹和第二子音频指纹(例如,在包括对第二音频指纹进行索引的步骤的实施例,是在索引之前进行切片);和/或,在获取音频指纹之前,预先对待识别音频以及已知音频按照预设的固定时间长度切片,得到多段时间长度相同的待识别音频片段和已知音频片段,然后分别获取各个待识别音频片段和已知音频片段的音频指纹,得到各个待识别音频片段的第一子音频指纹、各个已知音频片段的第二子音频指纹。之后,根据每个第一子音频指纹、第二子音频指纹进行前述的第一排名和第二排名的步骤,得到各个子音频指纹的识别结果,然后根据各个子音频指纹的识别结果确定原始的待识别音频的识别结果。

10 通过对音频或音频指纹按照固定长度切片的效果在于:1、使 TF-IDF 排名更加公平;2、求得的单体相似度、序列相似度评分更加准确;3、统一长度有利于音频指纹及指纹索引的存储。

在本公开的一些实施例中,第一音频指纹中的第一音频指纹单体以及第二音频指纹中的第二音频指纹单体在排列上具有时间性,例如,是按照时间的先后顺序排列的。这时,本公开的音频检索识别方法还包括:根据前述的相似度矩阵确定待识别音频与已知音频(具体地,可以是前述的识别结果中的音频)的重复片段。具体地,可以根据相似度矩阵中的直线的起点和终点得到两个音频中的重复片段的起止时间。

其中的根据相似度矩阵中的直线(例如匹配直线)来确定重复片段的具体方法可以是:根据直线的起点所对应的第一音频指纹单体的序数(或者说,相似度矩阵中的横坐标)确定待识别音频中的重复片段的开始时间,而根据该起点所对应的第二音频指纹单体的序数(或者说,相似度矩阵中的纵坐标)确定第一候选音频中的重复片段的开始时间;类似地,根据直线的终点的横坐标确定待识别音频中的重复片段的结束时间,而根据该终点的纵坐标确定第一候选音频中的重复片段的结束时间。

在本公开的一些实施例中(例如前述的图6和图7所示的实施例),步骤 S44 还包括:检测所得到的第一匹配直线或第二匹配直线的开头部分和结尾部分,判断该第一匹配直线/第二匹配直线的开头部分和结尾部分的点(单体相似度)是否达到预设的单体相似度设定值,去掉第一匹配直线/第二匹配直线的开头和结尾的未达到该单体相似度设定值(即单体相似度不高)的部分,保留中间一段直线并定义为第三匹配直线;根据该第三匹配直线的直线相似度来确定序列相似度评分,和/或根据该第三匹配直线的起点和终点确定已知音频与待识别音频的重复片段的起止时间。通过去掉匹配直线开头结尾的相似度不高的部分、保留中间一段相似度较高的直线之后,再确定已知音频与待识别音频的相似情况,能够提高音频检索识别的准确性,能够得到更准确的重复片段。

其中的去掉匹配直线开头/结尾的未达到该单体相似度设定值的部分的具体方法可以是：从匹配直线的起点/终点向中间依次检查，判断是否达到该单体相似度设定值，在找到第一个达到该单体相似度设定值的点后，去掉该点到起点/终点之间的多个点。

5 需要注意的是，该单体相似度设定值可以是一个单体相似度的具体数值，在检查时判断一个点是否达到该数值；也可以是一个比例值，在检查时判断一个点与第一匹配直线/第二匹配直线所包含的所有点的平均值或最大值相比，是否达到该比例值。

10 图 9 为本公开的音频检索识别装置 1000 一个实施例的示意性结构框图。请参阅图 9，本公开示例的音频检索识别装置 1000 主要包括：

音频指纹获取系统 1100，用于获取待识别音频（Query 音频）的音频指纹。其中，该音频指纹包括用于表示待识别音频的内容特征的第一部分以及用于表示该第一部分的可信程度的第二部分。

15 检索识别系统 1200，用于根据待识别音频的该音频指纹对待识别音频进行识别，得到识别结果。

20 图 10 为本公开一个实施例提供的音频指纹获取系统 1100 的示意性结构框图。请参阅图 10，本公开示例的音频指纹获取系统 1100 主要包括：声谱图转换模块 1101、特征点确定模块 1102、掩模确定模块 1103、均值能量确定模块 1104、音频指纹比特确定模块 1105、强弱权重比特确定模块 1106 以及音频指纹确定模块 1107。

其中，该声谱图转换模块 1101 用于将音频转换成声谱图（Spectrogram）。具体地，声谱图转换模块 1101 可具体用于通过短时傅里叶变换（Fast Fourier Transformation）将音频信号转换成时间-频率声谱图。

25 在本公开的实施例中，声谱图转换模块 1101 可包括梅尔变换子模块，用于利用梅尔（MEL）变换对声谱图进行预处理，利用梅尔变换能够将频谱分成多个频率区块（频率 bin），而所分成的频率区块的数目是可以配置的。另外，声谱图转换模块 1101 还可以包括人类听觉系统滤波子模块，用于对声谱图进行人类听觉系统滤波（Human Auditory System filtering），利用人类听觉系统滤波等非线性变换，能够使得声谱图中的频谱分布更适合人耳感知。

该特征点确定模块 1102 用于确定声谱图中的特征点。

35 具体地，该特征点确定模块 1102 可以具体用于采用多种标准中的一种来确定特征点，例如，可以将特征点选为声谱图中的能量的极大值点，或者，也可以选为能量的极小值点。

在本公开的实施例中，该特征点确定模块 1102 也可以不选能量的极值

点作为特征点，而是用于选取固定点作为特征点，例如可以选取频率值与预设的频率设定值相等的点（频率固定的点）。进一步地，该特征点确定模块 1102 可用于按照频率大小预设低频、中频、高频的多个频率设定值。

该掩模确定模块 1103 用于在声谱图上，在特征点的附近，为特征点确定一个或多个掩模（mask），每个掩模包含多个谱区域。具体地，在声谱图中，每个掩模所包含的多块谱区域可以是对称分布的。

该均值能量确定模块 1104，用于确定每个谱区域的均值能量。

该音频指纹比特确定模块 1105，用于根据掩模中的多块谱区域的均值能量，确定音频指纹比特（bit）。需要注意的是，该音频指纹比特即为前述的音频指纹中的用于表示音频的内容特征的第一部分。

在本公开实施例中，该音频指纹比特确定模块 1105 可具体用于根据一个掩模所包含的多个谱区域的均值能量的差值确定一个音频指纹比特。

该强弱权重比特确定模块 1106，用于判断音频指纹比特的可信程度，以确定每个音频指纹比特对应的强弱权重比特。需要注意的是，该强弱权重比特即为前述的音频指纹中的用于表示第一部分的可信程度的第二部分。

在本公开的实施例中，如果音频指纹比特是根据一个掩模所包含的多个谱区域均值能量的差值确定的，则该强弱权重比特确定模块 1106 具体用于：判断生成该音频指纹比特所使用的该差值的绝对值是否达到（或超过）预设的强弱比特阈值；如果达到强弱比特阈值，则将该音频指纹比特确定为强比特，并得到一个与该音频指纹比特对应的取值为 1 的强弱权重比特；如果未达到强弱比特阈值，则将该音频指纹比特确定为弱比特，并得到一个与该音频指纹比特对应的取值为 0 的强弱权重比特。

该音频指纹确定模块 1107，用于根据该音频指纹比特和该强弱权重比特确定音频的音频指纹。

本公开通过在提取音频指纹比特的同时，提取该音频指纹比特对应的强弱权重比特，能够为一段音频生成一个准确性高、鲁棒性好的音频指纹。

可选地，本公开的音频指纹获取系统 1100 还包括时间戳添加模块（图中未示出），用于为音频指纹添加一个时间戳字段，用于表示音频起始位置与该特征点的时间差的字段，该字段可以是一个 hash 值。而如果将特征点设为固定点，则可以不必包含本模块，即不必记录该时间戳。

可选地，本公开的音频指纹获取系统 1100 还包括音频标识添加模块（图中未示出），用于为音频指纹添加一个音频标识字段，用于记录该音频指纹所对应的音频信号的 ID 标识信息，该字段可以是一个 hash 值。

可选地，本公开的音频指纹获取系统 1100 还包括音频分割模块（图中未示出）和音频指纹组合模块（图中未示出）。该音频分割模块用于将原始

音频按时间分成多段子音频。利用前述的音频指纹获取系统 1100 所包含的模块，对各段子音频提取音频指纹，得到多个音频指纹。而音频指纹组合模块用于将提取的各段子音频的音频指纹组合在一起，得到该整段音频的音频指纹。

5 为了便于叙述和理解，不妨将待识别音频的音频指纹称为第一音频指纹，第一音频指纹所包含的音频指纹单体称为第一音频指纹单体，第一音频指纹单体对应的强弱权重单体称为第一强弱权重单体。

图 11 为本公开一个实施例提供的检索识别系统 1200 的示意性结构框图。请参阅图 11，本公开示例的检索识别系统 1200 主要包括：

10 第一排名模块 1210，用于根据该第一音频指纹，对多个已知音频进行第一排名，根据该第一排名的结果，取出前 k 个已知音频作为第一候选音频集合。其中的 k 为正整数，而 k 的具体取值是可以设置的。具体地，该第一排名模块 1210 用于根据每个单独的第一音频指纹单体与已知音频的匹配情况进行的排名。进一步地，该第一排名模块 1210 可以用于根据各个第
15 一音频指纹单体对已知音频进行的词频-逆向文件频率 TF-IDF 排名。

第二排名模块 1220，用于根据该第一音频指纹，对该第一候选音频集合进行第二排名，根据该第二排名的结果，取出第一候选音频集合中的前 n 个第一候选音频作为识别结果。其中的 n 为正整数，而 n 的具体取值是可以设置的。具体地，该第二排名模块 1220 用于根据多个顺序排列的第一音频指纹单体，对该第一候选音频集合中的音频进行的排名。
20

另外，检索识别系统 1200 还可用于根据该识别结果在 Meta 数据库中进行检索，能够得到该识别结果的音频信息，例如识别出的音频的名称、作者、出处等等。当识别结果包括多个音频时，可以同时提供多个识别出的音频的信息。

25 在本实施例中，第一排名模块 1210 在进行第一排名和/或第二排名模块 1220 在进行第二排名的过程中，在利用到音频指纹时，可以根据音频指纹中的强弱权重单体的对音频指纹单体进行加权。

前述的已知音频可以是一个音频数据库中的音频。在该音频数据库中存储有已知音频的音频指纹，并且在存储的已知音频的音频指纹中包含有
30 与第一音频指纹利用相同提取方法得到的与第一音频指纹相同类型的音频指纹，从而已知音频的音频指纹中也包括用于表示音频的内容特征的第一部分以及用于表示该第一部分的可信程度的第二部分。

在本公开的一些实施例中，本公开的音频检索识别装置 1000 还包括指纹索引获取模块（图中未示出），用于获取多个已知音频的音频指纹，为了
35 便于叙述和理解，不妨将已知音频的音频指纹称为第二音频指纹，第二音频指纹所包含的音频指纹单体称为第二音频指纹单体，第二音频指纹所包

含的强弱权重单体称为第二强弱权重单体；对该第二音频指纹进行索引，以预先得到已知音频的指纹索引。而第一排名模块 1210 具体用于将该指纹索引与待识别音频的第一音频指纹单体进行匹配，以对多个已知音频进行 TF-IDF 排名。

5 进一步地，该指纹索引获取模块可以用于获取已知音频的音频指纹的正排指纹索引 (forward index) 和倒排指纹索引 (inverted index)。

值得注意的是，该指纹索引获取模块可以用于根据强弱权重对第二音频指纹进行索引，以提高鲁棒性。

另外，如果对音频数据库中的所有已知音频都进行第一排名，可能会影响检索识别的效率，因此本公开的第一排名模块 1210 可以包括绝对匹配子模块 1211，用于在第一排名之前，先对音频数据库中的已知音频进行绝对匹配 (exact match)。

图 12 为本公开一个实施例提供的第一排名模块 1210 的示意性结构图。请参阅图 12，在本公开的一个实施例中，该第一排名模块 1210 具体包括：

15 绝对匹配子模块 1211，用于根据倒排指纹索引，统计各个第一音频指纹单体在哪些已知音频的第二音频指纹中出现，以从音频数据库中匹配出包含预设数量以上第一音频指纹单体的已知音频作为第二候选音频集合。值得注意的是，该绝对匹配子模块 1211 可以具体用于根据一个第一音频指纹单体对应的强弱权重单体，仅判断该第一音频指纹单体中的强比特在已知音频的第二音频指纹中的匹配情况，而忽略该第一音频指纹单体中的弱比特的匹配情况，以提高鲁棒性。

词频确定子模块 1212，用于基于正排指纹索引，确定一个第一音频指纹单体在一个第二候选音频的第二音频指纹中的词频。值得注意的是，该正排指纹索引可以是前述的根据强弱权重得到的指纹索引。

25 文件频率确定子模块 1213，用于基于倒排指纹索引，确定一个第一音频指纹单体的文件频率。值得注意的是，该倒排指纹索引可以是前述的根据强弱权重得到的指纹索引。

词频-逆向文件频率评分子模块 1214，用于根据各个第一音频指纹单体在一个第二候选音频的第二音频指纹中的词频以及各个第一音频指纹单体的文件频率，确定该第二候选音频的词频-逆向文件频率评分。

30 第一排名子模块 1215，用于根据得到的各个第二候选音频的词频-逆向文件频率评分对第二候选音频集合进行排名，得到第一排名的结果，从该第一排名结果中取出前 k 个第二候选音频作为第一候选音频集合；该第一排名子模块 1215 还可用于将各个第一候选音频的第二音频指纹 (正排指纹索引) 返回给第二排名模块 1220，以备在后续的进一步处理。

在本公开的一些实施例中，该第二排名为根据多个顺序排列的第一音

频指纹单体所组成的具有先后顺序的序列在第一候选音频的音频指纹中出现的情况，对该第一候选音频集合中的音频进行的排名。具体地，该第二排名模块 1220 用于：根据已知音频的指纹索引与第一音频指纹得到该第一候选音频集合中的音频的相似度矩阵，根据该相似度矩阵对该第一候选音频集合中的音频进行排名。值得注意的是，该第二排名模块 1220 可以具体用于：在确定相似度矩阵的过程中，根据第一音频指纹对应的强弱权重和/或已知音频的指纹索引中的强弱权重进行加权，并利用加权后的相似度矩阵对第一候选音频集合中的音频进行排名，以提高鲁棒性。

图 13 为本公开一个实施例提供的第二排名模块 1220 的示意性结构图。请参阅图 13，在本公开的一个实施例中，该第二排名模块 1220 具体包括：

第二音频指纹获取子模块 1221，用于获取第一候选音频集合中的一个第一候选音频（事实上每个第一候选音频都是已知音频）的第二音频指纹。具体地，可以根据已知音频的指纹索引（例如，正排指纹索引）获取该第二音频指纹。在本公开的一些示例中，第一音频指纹中包含有与各个第一音频指纹单体对应的强弱权重单体（不妨称之为第一强弱权重单体），和/或第二音频指纹中包含有与各个第二音频指纹单体对应的强弱权重单体（不妨称之为第二强弱权重单体）。

单体相似度第一确定子模块 1222，用于确定该第一候选音频的第二音频指纹所包含的各个第二音频指纹单体与各个第一音频指纹单体之间的单体相似度。值得注意的是，单体相似度第一确定子模块 1222 可以具体用于：在确定该单体相似度的过程中，根据第一强弱权重单体和/或第二强弱权重单体，对各个第一音频指纹单体、第二音频指纹单体进行加权，然后根据加权后的第一、第二音频指纹单体确定该单体相似度。在本公开的一种示例中，由于音频数据库中的数据信息的准确性更高，可以利用第二强弱权重单体分别对第一音频指纹单体、第二音频指纹单体进行加权。

相似度矩阵第一确定子模块 1223，用于根据各个单体相似度，确定该第一候选音频与待识别音频之间的相似度矩阵。

序列相似度评分确定子模块 1224，用于根据一个第一候选音频的相似度矩阵，确定该第一候选音频的序列相似度评分。具体地，该序列相似度评分确定子模块 1224 具体用于根据相似度矩阵中的直线来确定该的序列相似度评分。

第二排名子模块 1225，用于根据各个第一候选音频的该序列相似度评分对第一候选音频集合进行排名，得到第二排名的结果，从该第二排名结果中取出前 n 个第一候选音频作为识别结果。

在本公开的一个实施例中，该序列相似度评分确定子模块 1224 具体用于利用前述的匀速音频法的各个具体步骤来确定该序列相似度评分。

在本公开的一个实施例中，该序列相似度评分确定子模块 1224 具体用于利用前述的动态规划法的各个具体步骤来确定该序列相似度评分。

进一步地，其中相似度矩阵可以是由多种音频相似度综合考量得到的。图 14 为本公开一个实施例的基于多种类型的第一音频指纹和第二音频指纹确定相似度矩阵的音频检索识别装置 1000 的示意性结构框图。请参阅图 14，在本公开的一个实施例中，本公开的音频检索识别装置 1000 包括：

多类型第一音频指纹获取模块 1300，用于利用多种音频指纹获取方法，获取待识别音频的多种类型的第一音频指纹，每种类型的第一音频指纹包含多个用于表示音频内容特征的第一部分不妨称为第一音频指纹单体。可选地，至少一些类型的第一音频指纹包含用于表示第一部分的可信程度的第二部分。

多类型第二音频指纹获取模块 1400，用于获取一个已知音频（具体地，可以是前述的第一候选音频集合中的音频）的多种类型的第二音频指纹，每种类型的第二音频指纹包含多个用于表示音频内容特征的第一部分不妨称为第二音频指纹单体。可选地，至少一些类型的第二音频指纹包含用于表示第一部分的可信程度的第二部分。

单体相似度第二确定子模块 1500，用于分别确定同种类型的该第二音频指纹单体与该第一音频指纹单体之间的单体相似度。从而对应于多种类型的音频指纹，能够得到一个已知音频的多种单体相似度。

相似度矩阵第二确定子模块 1600，用于确定多种单体相似度的平均值或最小值，并根据多种单体相似度的该平均值或该最小值确定该已知音频的相似度矩阵。

进而前述的序列相似度评分确定子模块 1224 用于根据该基于多种单体相似度的平均值或最小值的相似度矩阵来确定序列相似度评分。

在本公开的一些实施例中，该音频检索识别装置 1000 还包括音频切片模块（图中未示出）。该音频切片模块用于在进行第一排名之前，对获取的待识别音频的第一音频指纹以及已知音频的第二音频指纹按照预设的固定长度切片，得到多个长度相同（包含相同数量的音频指纹单体）的第一子音频指纹和第二子音频指纹；和/或，该音频切片模块用于在获取音频指纹之前，预先对待识别音频以及已知音频按照预设的固定时间长度切片，得到多段时间长度相同的待识别音频片段和已知音频片段，然后分别获取各个待识别音频片段和已知音频片段的音频指纹，得到各个待识别音频片段的第一子音频指纹、各个已知音频片段的第二子音频指纹。而前述的第一排名模块 1210 和第二排名模块 1220 用于根据每个第一子音频指纹、第二子音频指纹进行前述的第一排名和第二排名的步骤，得到各个子音频指纹的识别结果，然后根据各个子音频指纹的识别结果确定原始的待识别音频

的识别结果。

在本公开的一些实施例中，第一音频指纹中的第一音频指纹单体以及第二音频指纹中的第二音频指纹单体在排列上具有时间性。这时，本公开的音频检索识别装置 1000 还包括重复音频片段确定模块（图中未示出），
5 用于根据前述的相似度矩阵确定待识别音频与已知音频的重复片段。具体地，该重复媒体片段确定模块具体用于根据相似度矩阵中的直线的起点和终点得到两个音频中的重复片段的起止时间。

图 15 是图示根据本公开的实施例的音频检索识别硬件装置的硬件框图。如图 15 所示，根据本公开实施例的音频检索识别硬件装置 2000 包括
10 存储器 2001 和处理器 2002。音频检索识别硬件装置 2000 中的各组件通过总线系统和/或其它形式的连接机构（未示出）互连。

该存储器 2001 用于存储非暂时性计算机可读指令。具体地，存储器 2001 可以包括一个或多个计算机程序产品，该计算机程序产品可以包括各种形式的计算机可读存储介质，例如易失性存储器和/或非易失性存储器。
15 该易失性存储器例如可以包括随机存取存储器（RAM）和/或高速缓冲存储器（cache）等。该非易失性存储器例如可以包括只读存储器（ROM）、硬盘、闪存等。

该处理器 2002 可以是中央处理单元（CPU）或者具有数据处理能力和/或指令执行能力的其它形式的处理单元，并且可以控制音频检索识别硬件
20 装置 2000 中的其它组件以执行期望的功能。在本公开的一个实施例中，该处理器 2002 用于运行该存储器 2001 中存储的该计算机可读指令，使得该音频检索识别硬件装置 2000 执行前述的本公开各实施例的音频检索识别方法的全部或部分步骤。

图 16 是图示根据本公开的实施例的计算机可读存储介质的示意图。如图
25 图 16 所示，根据本公开实施例的计算机可读存储介质 3000，其上存储有非暂时性计算机可读指令 3001。当该非暂时性计算机可读指令 3001 由处理器运行时，执行前述的本公开各实施例的音频检索识别方法的全部或部分步骤。

图 17 是图示根据本公开实施例的终端设备的硬件结构示意图。终端设备
30 可以以各种形式来实施，本公开中的终端设备可以包括但不限于诸如移动电话、智能电话、笔记本电脑、数字广播接收器、PDA（个人数字助理）、PAD（平板电脑）、PMP（便携式多媒体播放器）、导航装置、车载终端设备、车载显示终端、车载电子后视镜等等的移动终端设备以及诸如数字 TV、台式计算机等等的固定终端设备。

如图 17 所示，终端设备 4100 可以包括无线通信单元 4110、A/V（音频/视频）输入单元 4120、用户输入单元 4130、感测单元 4140、输出单元 4150、
35

存储器 4160、接口单元 4170、控制器 4180 和电源单元 4190 等等。图 17 示出了具有各种组件的终端设备，但是应理解的是，并不要求实施所有示出的组件。可以替代地实施更多或更少的组件。

其中，无线通信单元 4110 允许终端设备 4100 与无线通信系统或网络之间的无线电通信。A/V 输入单元 4120 用于接收音频或视频信号。用户输入单元 4130 可以根据用户输入的命令生成键输入数据以控制终端设备的各种操作。感测单元 4140 检测终端设备 4100 的当前状态、终端设备 4100 的位置、用户对于终端设备 4100 的触摸输入的有无、终端设备 4100 的取向、终端设备 4100 的加速或减速移动和方向等等，并且生成用于控制终端设备 4100 的操作的命令或信号。接口单元 4170 用作至少一个外部装置与终端设备 4100 连接可以通过的接口。输出单元 4150 被构造为以视觉、音频和/或触觉方式提供输出信号。存储器 4160 可以存储由控制器 4180 执行的处理和控制操作的软件程序等等，或者可以暂时地存储已经输出或将要输出的数据。存储器 4160 可以包括至少一种类型的存储介质。而且，终端设备 4100 可以与通过网络连接执行存储器 4160 的存储功能的网络存储装置协作。控制器 4180 通常控制终端设备的总体操作。另外，控制器 4180 可以包括用于再现或回放多媒体数据的多媒体模块。控制器 4180 可以执行模式识别处理，以将在触摸屏上执行的手写输入或者图片绘制输入识别为字符或图像。电源单元 4190 在控制器 4180 的控制下接收外部电力或内部电力并且提供操作各元件和组件所需的适当的电力。

本公开提出的音频检索识别方法的各种实施方式可以以使用例如计算机软件、硬件或其任何组合的计算机可读介质来实施。对于硬件实施，本公开提出的音频检索识别方法的各种实施方式可以通过使用特定用途集成电路 (ASIC)、数字信号处理器 (DSP)、数字信号处理装置 (DSPD)、可编程逻辑装置 (PLD)、现场可编程门阵列 (FPGA)、处理器、控制器、微控制器、微处理器、被设计为执行这里描述的功能的电子单元中的至少一种来实施，在一些情况下，本公开提出的音频检索识别方法的各种实施方式可以在控制器 4180 中实施。对于软件实施，本公开提出的音频检索识别方法的各种实施方式可以与允许执行至少一种功能或操作的单独的软件模块来实施。软件代码可以由以任何适当的编程语言编写的软件应用程序（或程序）来实施，软件代码可以存储在存储器 4160 中并且由控制器 4180 执行。

以上，根据本公开实施例的音频检索识别方法、装置、硬件装置、计算机可读存储介质以及终端设备，通过获取并利用音频对象的包括用于表示音频内容特征的第一部分和用于表示第一部分的可信程度的第二部分的音频指纹特征来进行音频检索识别，能够大大提高音频检索识别的准确性、鲁棒性和效率。

以上结合具体实施例描述了本公开的基本原理，但是，需要指出的是，在本公开中提及的优点、优势、效果等仅是示例而非限制，不能认为这些优点、优势、效果等是本公开的各个实施例必须具备的。另外，上述公开的具体细节仅是为了示例的作用和便于理解的作用，而非限制，上述细节并不限制本公开为必须采用上述具体的细节来实现。

本公开中涉及的器件、装置、设备、系统的方框图仅作为例示性的例子并且不意图要求或暗示必须按照方框图示出的方式进行连接、布置、配置。如本领域技术人员将认识到的，可以按任意方式连接、布置、配置这些器件、装置、设备、系统。诸如“包括”、“包含”、“具有”等等的词语是开放性词汇，指“包括但不限于”，且可与其互换使用。这里所使用的词汇“或”和“和”指词汇“和/或”，且可与其互换使用，除非上下文明确指示不是如此。这里所使用的词汇“诸如”指词组“诸如但不限于”，且可与其互换使用。

另外，如在此使用的，在以“至少一个”开始的项的列举中使用的“或”指示分离的列举，以便例如“A、B 或 C 的至少一个”的列举意味着 A 或 B 或 C，或 AB 或 AC 或 BC，或 ABC（即 A 和 B 和 C）。此外，措辞“示例的”不意味着描述的例子是优选的或者比其他例子更好。

还需要指出的是，在本公开的系统和方法中，各部件或各步骤是可以分解和/或重新组合的。这些分解和/或重新组合应视为本公开的等效方案。

可以不脱离由所附权利要求定义的教导的技术而进行对在此所述的技术的各种改变、替换和更改。此外，本公开的权利要求的范围不限于以上所述的处理、机器、制造、事件的组成、手段、方法和动作的具体方面。可以利用与在此所述的相应方面进行基本相同的功能或者实现基本相同的结果的当前存在的或者稍后要开发的处理、机器、制造、事件的组成、手段、方法或动作。因而，所附权利要求包括在其范围内的这样的处理、机器、制造、事件的组成、手段、方法或动作。

提供所公开的方面的以上描述以使本领域的任何技术人员能够做出或者使用本公开。对这些方面的各种修改对于本领域技术人员而言是非常显而易见的，并且在此定义的一般原理可以应用于其他方面而不脱离本公开的范围。因此，本公开不意图被限制到在此示出的方面，而是按照与在此公开的原理和新颖的特征一致的最宽范围。

为了示例和描述的目的已经给出了以上描述。此外，此描述不意图将本公开的实施例限制到在此公开的形式。尽管以上已经讨论了多个示例方面和实施例，但是本领域技术人员将认识到其某些变型、修改、改变、添加和子组合。