(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2024 年 3 月 14 日 (14.03.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/051481 A1

(51) 国际专利分类号:
G06F 16/65 (2019.01) *G10L 25/27* (2013.01)

(21) 国际申请号: PCT/CN2023/114040

(22) 国际申请日: 2023 年 8 月 21 日 (21.08.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202211088204.6 2022年9月7日 (07.09.2022) CN

(71) 申请人: 腾 讯 科 技 (深 圳) 有 限 公 司 **(TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED)** [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(72) 发明人: 朱鸿宁 **(ZHU, Hongning)**; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(74) 代理人:北京三高永信知识产权代理有限责任公司**(BEIJING SAN GAO YONG XIN INTELLECTUAL PROPERTY AGENCY CO., LTD.)**; 中国北京市海淀区上地信息产业基地三街1号楼四层C段457, Beijing 100085 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA,

WO 2024/051481 A1

(54) **Title:** AUDIO PROCESSING METHOD AND APPARATUS, DEVICE, READABLE STORAGE MEDIUM, AND PROGRAM PRODUCT

(54) 发明名称: 音频处理方法、装置、设备、可读存储介质及程序产品

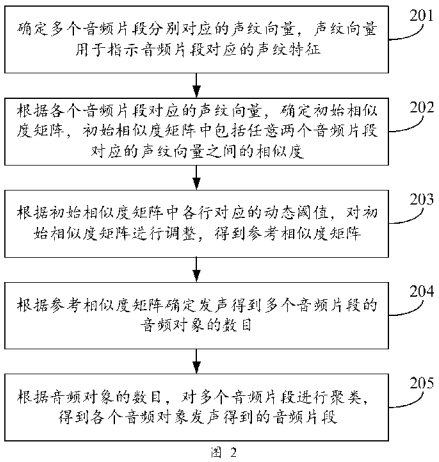

图 2

201 Determine voiceprint vectors respectively corresponding to a plurality of audio clips, the voiceprint vectors being used for indicating voiceprint features corresponding to the audio clips

202 Determine an initial similarity matrix according to the voiceprint vectors corresponding to the audio clips, the initial similarity matrix comprising the similarity between the voiceprint vectors corresponding to any two audio clips

203 Adjust the initial similarity matrix according to dynamic thresholds corresponding to rows in the initial similarity matrix to obtain a reference similarity matrix

204 According to the reference similarity matrix, determine the number of audio objects of the plurality of audio clips obtained by sound production

205 Perform clustering on the plurality of audio clips according to the number of audio objects to obtain audio clips obtained by the audio objects by sound production

(57) **Abstract:** An audio processing method and apparatus, a device, a readable storage medium, and a program product, relating to the technical field of computers. The method comprises: determining voiceprint vectors respectively corresponding to a plurality of audio clips, the voiceprint vectors being used for indicating voiceprint features corresponding to the audio clips (201); determining an initial similarity matrix according to the voiceprint vectors corresponding to the audio clips, the initial similarity matrix comprising the similarity between the voiceprint vectors corresponding to any two audio clips (202); adjusting the initial similarity matrix according

[见续页]

PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

to dynamic thresholds corresponding to rows in the initial similarity matrix to obtain a reference similarity matrix (203); determining the number of audio objects in the plurality of audio clips according to the reference similarity matrix (204); and performing clustering on the plurality of audio clips according to the number of audio objects to obtain audio clips corresponding to the audio objects (205).

(57) 摘要: 一种音频处理方法、装置、设备、可读存储介质及程序产品，属于计算机技术领域。方法包括：确定多个音频片段分别对应的声纹向量，声纹向量用于指示音频片段对应的声纹特征(201)；根据各个音频片段对应的声纹向量，确定初始相似度矩阵，初始相似度矩阵中包括任意两个音频片段对应的声纹向量之间的相似度(202)；根据初始相似度矩阵中各行对应的动态阈值，对初始相似度矩阵进行调整，得到参考相似度矩阵(203)；根据参考相似度矩阵确定多个音频片段中存在的音频对象的数目(204)；根据音频对象的数目，对多个音频片段进行聚类，得到各个音频对象对应的音频片段(205)。

说 明 书

音频处理方法、装置、设备、可读存储介质及程序产品

本申请要求于2022年09月07日提交的申请号为202211088204.6、发明名称为“音频处理方法、装置、设备及计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请实施例涉及计算机技术领域，特别涉及一种音频处理方法、装置、设备、可读存储介质及程序产品。

背景技术

随着计算机技术的不断发展，音频处理方式也越来越多。例如，音频对象聚类是一种音频处理方式。音频对象聚类是确定多个音频片段中包括的音频对象的数目，以及各个音频对象对应的音频片段。

相关技术中，获取多个音频片段，确定各个音频片段对应的声纹向量，根据各个声纹向量之间的相似度，确定多个音频片段中存在的音频对象的数目，根据音频对象的数目对多个音频片段进行聚类，得到各个音频对象对应的音频片段。

然而，由于音频片段是通过录音设备获取到的，当录音设备、音频对象的语气、情绪不同时，会使得一个音频对象的音色产生偏差，且音频对象所处的环境也会对声纹向量的确定造成干扰，使得确定的声纹向量不够准确，导致音频对象聚类的准确率较低、音频处理效果较差。

发明内容

本申请实施例提供了一种音频处理方法、装置、设备、可读存储介质及程序产品，可以提高音频对象聚类的准确率。

一方面，本申请实施例提供了一种音频处理方法，由计算机设备执行，所述方法包括：

确定多个音频片段分别对应的声纹向量，所述声纹向量用于表示所述音频片段对应的声纹特征；

根据各个音频片段对应的声纹向量，确定初始相似度矩阵，所述初始相似度矩阵中包括任意两个音频片段对应的声纹向量之间的相似度；

根据所述初始相似度矩阵中各行对应的动态阈值，对所述初始相似度矩阵进行调整，得到参考相似度矩阵，所述动态阈值用于对不同相似度之间的相似度差值进行调节；

根据所述参考相似度矩阵确定发声得到所述多个音频片段的音频对象的数目；

根据所述音频对象的数目，对所述多个音频片段进行聚类，得到各个音频对象发声得到的音频片段。

另一方面，本申请实施例提供了一种音频处理装置，所述装置包括：

确定模块，用于确定多个音频片段分别对应的声纹向量，所述声纹向量用于表示所述音频片段对应的声纹特征；

所述确定模块，还用于根据各个音频片段对应的声纹向量，确定初始相似度矩阵，所述初始相似度矩阵中包括任意两个音频片段对应的声纹向量之间的相似度；

调整模块，用于根据所述初始相似度矩阵中各行对应的动态阈值，对所述初始相似度矩阵进行调整，得到参考相似度矩阵，所述动态阈值用于对不同相似度之间的相似度差值进行调节；

所述确定模块，还用于根据所述参考相似度矩阵确定发声得到所述多个音频片段的音频

对象的数目；

聚类模块，用于根据所述音频对象的数目，对所述多个音频片段进行聚类，得到各个音频对象发声得到的音频片段。

在一种可能的实现方式中，所述确定模块，还用于对于所述初始相似度矩阵中的任一行，按照第一顺序对所述任一行中位于预设相似度范围内的相似度进行排序，得到第一排序结果；确定所述第一排序结果中相邻的两个相似度之间的相似度差值，得到多个相似度差值；在所述多个相似度差值中确定满足第一要求的相似度差值；根据所述满足第一要求的相似度差值，确定所述任一行对应的动态阈值。

在一种可能的实现方式中，所述调整模块，用于将所述初始相似度矩阵第 k 行包括的相似度中，小于第 k 行对应的动态阈值的相似度调整为第一数值，并基于各行的调整结果得到所述参考相似度矩阵，k 为正整数；或者，将所述初始相似度矩阵第 k 行包括的相似度中，小于所述第 k 行对应的动态阈值的相似度与第二数值相乘，并基于各行的调整结果得到所述参考相似度矩阵。

在一种可能的实现方式中，所述确定模块，用于根据多个参考参数，对所述参考相似度矩阵进行处理，得到各个参考参数对应的相似度矩阵；根据所述多个参考参数和所述各个参考参数对应的相似度矩阵，确定所述多个音频片段中存在的音频对象的数目。

在一种可能的实现方式中，所述确定模块，用于对于所述多个参考参数中的任一参考参数，根据所述任一参考参数，对所述参考相似度矩阵进行数值调整，得到第一相似度矩阵，所述数值调整用于简化所述参考相似度矩阵；对所述第一相似度矩阵进行对称化处理，得到第二相似度矩阵，所述第二相似度矩阵中位于第 i 行第 j 列的相似度与位于第 j 行第 i 列的相似度相同，所述 i 和所述 j 为不大于所述多个音频片段的个数的正整数；对所述第二相似度矩阵进行行列扩散，得到第三相似度矩阵，所述第三相似度矩阵用于生成多个音频对象之间的边界；对所述第三相似度矩阵进行比例调整，得到第四相似度矩阵，所述比例调整用于将所述第三相似度矩阵中各行包括的相似度调整在同一个范围内；对所述第四相似度矩阵进行对称化处理，得到所述任一参考参数对应的相似度矩阵。

在一种可能的实现方式中，所述确定模块，用于对于所述参考相似度矩阵各行包括的多个相似度，将任一参考参数个满足第三要求的相似度之外的相似度调整为第三数值，得到所述第一相似度矩阵；或者，将所述参考相似度矩阵包括的多个相似度中，除任一参考参数个满足第三要求的相似度之外的相似度与第四数值相乘，得到所述第一相似度矩阵。

在一种可能的实现方式中，所述确定模块，用于确定所述第一相似度矩阵对应的转置矩阵；将所述第一相似度矩阵和所述第一相似度矩阵对应的转置矩阵中位于相同位置的相似度相加，得到待调整相似度矩阵；对所述待调整相似度矩阵包括的多个相似度进行取半操作，得到所述第二相似度矩阵。

在一种可能的实现方式中，所述确定模块，用于确定所述第一相似度矩阵中位于所述第 i 行第 j 列的相似度，与所述第一相似度矩阵中位于所述第 j 行第 i 列的相似度中最大的相似度，将所述最大的相似度作为所述第二相似度矩阵中位于所述第 i 行第 j 列和所述第 j 行第 i 列的相似度，得到所述第二相似度矩阵。

在一种可能的实现方式中，所述确定模块，用于确定所述第二相似度矩阵对应的转置矩阵；根据所述第二相似度矩阵和所述第二相似度矩阵对应的转置矩阵，确定所述第三相似度矩阵，所述第三相似度矩阵中位于第 m 行第 n 列的相似度基于所述第二相似度矩阵中位于所述第 m 行的相似度和所述第二相似度矩阵对应的转置矩阵中位于所述第 n 列的相似度确定，所述 m、所述 n 为不大于所述多个音频片段的个数的正整数。

在一种可能的实现方式中，所述确定模块，用于根据所述第三相似度矩阵中各行包括的多个相似度，确定各行对应的最大相似度；将所述第三相似度矩阵中各行包括的多个相似度分别与所述各行对应的最大相似度相除，得到所述第四相似度矩阵。

在一种可能的实现方式中，所述确定模块，用于根据所述多个参考参数和所述各个参考

参数对应的相似度矩阵，确定所述各个参考参数对应的比例值，所述比例值用于指示所述参考参数对应的相似度矩阵中保留的相似度的数量；根据所述各个参考参数对应的比例值，确定所述多个音频片段中存在的音频对象的数目。

在一种可能的实现方式中，所述确定模块，用于对于所述多个参考参数中的任一参考参数，对所述任一参考参数对应的相似度矩阵进行拉普拉斯变换，得到所述任一参考参数对应的拉普拉斯矩阵；对所述拉普拉斯矩阵进行奇异值分解，得到多个参考特征值；在所述多个参考特征值中确定第二特征值和第一数量个第一特征值，所述第二特征值为所述多个参考特征值中的最大值，所述第一特征值为按照第二顺序对所述多个参考特征值进行排序后满足第二要求的参考特征值；确定所述第一数量个第一特征值中相邻的两个第一特征值之间的差值，得到多个特征值差值；根据所述第二特征值，对第一特征值差值进行归一化处理，得到归一化之后的特征值差值，所述第一特征值差值为所述多个特征值差值中最大的特征值差值；根据所述归一化之后的特征值差值和所述任一参考参数，确定所述任一参考参数对应的比例值。

在一种可能的实现方式中，所述确定模块，用于根据所述各个参考参数对应的比例值，在所述多个参考参数中确定第一参数，所述第一参数为所述多个参考参数中对应的比例值最小的参考参数；确定所述第一参数对应的多个特征值差值；调用第一函数对所述第一参数对应的多个特征值差值进行处理，得到所述多个音频片段中存在的音频对象的数目。

在一种可能的实现方式中，所述聚类模块，用于对所述第一参数对应的相似度矩阵进行奇异值分解，得到多个分解特征值；在所述多个分解特征值中确定与所述音频对象的数目对应分解特征值；确定所述音频对象的数目个分解特征值分别对应的特征向量，并生成分解矩阵，所述分解矩阵的行数为所述音频对象的数目，列数为所述音频片段的数目；根据所述分解矩阵，确定所述多个音频片段分别对应的特征向量，所述特征向量用于指示对应的音频片段；根据所述音频对象的数目和所述多个音频片段分别对应的特征向量，对所述多个音频片段进行聚类，各个音频对象发声得到的音频片段。

另一方面，本申请实施例提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条程序代码，所述至少一条程序代码由所述处理器加载并执行，以使计算机设备实现上述任一所述的音频处理方法。

另一方面，还提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有至少一条程序代码，所述至少一条程序代码由处理器加载并执行，以使计算机实现上述任一所述的音频处理方法。

另一方面，还提供了一种计算机程序或计算机程序产品，所述计算机程序或计算机程序产品中存储有至少一条计算机指令，所述至少一条计算机指令由处理器加载并执行，以使计算机实现上述任一种音频处理方法。

本申请实施例提供的技术方案至少带来如下有益效果。

根据初始相似度矩阵中各行对应的动态阈值，对初始相似度矩阵进行调整，进而得到参考相似度矩阵，通过动态阈值调整过程，能够拉近同一音频对象的音频片段的声纹向量的相似度，拉远不同音频对象的音频片段的声纹向量的相似度，使得根据参考相似度矩阵，确定的音频对象的数目更加准确；再根据准确率较高的音频对象的数目，对多个音频片段进行聚类，得到各个音频对象对应的音频片段，使得确定的各个音频对象对应的音频片段的准确性较高，音频对象聚类的准确率较高，进而能够提高音频片段的音频处理效果。

附图说明

图 1 是本申请实施例提供的一种音频处理方法的实施环境示意图；

图 2 是本申请实施例提供的一种音频处理方法的流程图；

图 3 是本申请实施例提供的一种相似度矩阵的确定过程的示意图；

图 4 是本申请实施例提供的另一种音频处理方法的流程图；

图 5 是本申请实施例提供的一种音频处理装置的结构示意图；

图 6 是本申请实施例提供的一种终端设备的结构示意图；

图 7 是本申请实施例提供的一种服务器的结构示意图。

具体实施方式

在示例性实施例中，本申请实施例提供的音频处理方法可应用于各种场景，包括但不限于云技术、人工智能、智慧交通、辅助驾驶、游戏等。

人工智能（Artificial Intelligence，AI）是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，人工智能企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。

本申请实施例提供的方案涉及人工智能技术中的机器学习技术，机器学习（Machine Learning，ML）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、示教学习等技术。

随着人工智能技术研究和进步，人工智能技术在多个领域展开研究和应用，例如常见的智能家居、智能穿戴设备、虚拟助理、智能音箱、智能营销、无人驾驶、自动驾驶、无人机、机器人、智能医疗、智能客服、车联网、自动驾驶、智慧交通等。

图 1 是本申请实施例提供的一种音频处理方法的实施环境示意图，如图 1 所示，该实施环境包括：终端设备 101 和服务器 102。

本申请实施例提供的音频处理方法可以由终端设备 101 执行，也可以由服务器 102 执行，还可以由终端设备 101 和服务器 102 共同执行，本申请实施例对此不进行限定。对于本申请实施例提供的音频处理方法由终端设备 101 和服务器 102 共同执行的情况，服务器 102 承担主要计算工作，终端设备 101 承担次要计算工作；或者，服务器 102 承担次要计算工作，终端设备 101 承担主要计算工作；或者，服务器 102 和终端设备 101 二者之间采用分布式计算架构进行协同计算。

可选地，终端设备 101 可以是任何一种可与用户通过键盘、触摸板、触摸屏、遥控器、语音交互或手写设备等一种或多种方式进行人机交互的电子产品。终端设备 101 包括但不限于手机、电脑、智能语音交互设备、智能家电、车载终端、飞行器等。服务器 102 为一台服务器，或者为多台服务器组成的服务器集群，或者为云计算平台和虚拟化中心，或者为区块链系统中的节点中的任意一种，本申请实施例对此不加以限定。服务器 102 与终端设备 101 通过有线网络或无线网络进行通信连接。服务器 102 具有数据接收功能、数据处理功能和数据发送功能。当然，服务器 102 还可以具有其他功能，本申请实施例对此不加以限定。

本领域技术人员应能理解上述终端设备 101 和服务器 102 仅为举例说明，其他现有的或者今后可能出现的终端设备或服务器，如可适用于本申请，也应包含在本申请的保护范围之内，并在此以引用方式包含于此。

本申请实施例提供了一种音频处理方法，该方法由计算机设备执行，该方法可应用于上述图 1 所示的实施环境，计算机设备可以是图 1 中的终端设备 101，也可以是图 1 中的服务器 102，本申请实施例对此不进行限定。以图 2 所示的本申请实施例提供的一种音频处理方法的流程图为例，如图 2 所示，该方法包括下述步骤 201 至步骤 205。

在步骤 201 中，确定多个音频片段分别对应的声纹向量，声纹向量用于指示音频片段对应的声纹特征。

在本申请实施例中，在确定多个音频片段分别对应的声纹向量之前，需要先获取多个音频片段，多个音频片段为至少两个音频片段，每个音频片段对应于一个音频对象，音频对象为音频片段的发声对象，其中，不同音频片段对应的音频对象相同或者不同。本申请实施例对多个音频片段的获取过程不进行限定。示例性地，计算机设备的存储空间中存储有多个候选片段，从多个候选片段中获取多个音频片段。

可选地，还可以获取一个语音数据，语音数据中包括多个音频对象的音频片段，对语音数据进行分割处理，得到多个语音片段，在多个语音片段中确定待处理的多个音频片段。例如，将多个语音片段作为待处理的音频片段，或者在多个语音片段中选取一部分语音片段作为待处理的音频片段。示意性的，获取歌曲数据作为语音数据，根据歌词分段情况将语音数据分割为多个语音片段作为音频片段。

需要说明的是，多个音频片段分别对应的时长可以相同，也可以不同，本申请实施例对此不进行限定。例如，多个音频片段分别对应的时长均为 2 秒，又例如，多个音频片段中有的音频片段对应的时长为 2 秒，有的音频片段对应的时长为 5 秒。

可选地，确定出待处理的多个音频片段之后，对各个音频片段进行特征提取，得到各个音频片段对应的特征。根据各个音频片段对应的特征，确定各个音频片段对应的声纹向量。示例性地，各个音频片段对应的特征可以是各个音频片段对应的 MFCC（Mel-scaleFrequency Cepstral Coefficients，梅尔倒谱系数），也可以是各个音频片段对应的梅尔频谱特征，还可以是其他特征，本申请实施例对此不进行限定。

在一种可能的实现方式中，根据各个音频片段对应的特征，确定各个音频片段对应的声纹向量的过程包括：将音频片段对应的特征输入声纹提取模型，将声纹提取模型输出的结果作为音频片段对应的声纹向量。可选地，声纹提取模型可以是任意一种模型，本申请实施例对此不进行限定。示例性地，声纹提取模型可以为 CLDNN（Convolution-Longshort-Term Mermony-Fully-Connected Deep Neural Networks，卷积-长短期记忆力-全连接神经网络）模型，也可以为基于 TDNN（Time Delay Neural Network，时延神经网络）的 X-vector（声纹识别领域主流的 baseline 模型框架），还可以是 ecapa-tdnn（一种提取语音全局特征的模型）。

在步骤 202 中，根据各个音频片段对应的声纹向量，确定初始相似度矩阵，初始相似度矩阵中包括任意两个音频片段对应的声纹向量之间的相似度。

在一种可能的实现方式中，在上述步骤 201 中确定出各个音频片段对应的声纹向量，根据各个音频片段对应的声纹向量，确定任意两个音频片段对应的声纹向量之间的相似度，得到初始相似度矩阵。

可选地，根据各个音频片段对应的声纹向量，按照下述公式（1）确定任意两个音频片段对应的声纹向量之间的相似度。

$$a_{i,j} = d(v_i, v_j) \quad i,j \in [1, N] \quad \text{公式（1）}$$

在上述公式（1）中，$a_{i,j}$ 为第 i 个音频片段对应的声纹向量和第 j 个音频片段对应的声纹向量之间的相似度，$d(v_i, v_j)$ 为距离公式，v_i 为第 i 个音频片段对应的声纹向量，v_j 为第 j 个音频片段对应的声纹向量，N 为多个音频片段的总数量。

可选地，可以将任意两个音频片段对应的声纹向量之间的余弦相似度距离作为任意两个音频片段对应的声纹向量之间的相似度。当然，还可以通过其他方式确定任意两个音频片段对应的声纹向量之间的相似度，本申请实施例对此不进行限定。

需要说明的是，初始相似度矩阵的行数为多个音频片段的个数，列数为多个音频片段的个数。初始相似度矩阵为对称化矩阵，也即是初始相似度矩阵中位于第 i 行第 j 列的相似度与位于第 j 行第 i 列的相似度是相同的。其中，i，j 均为不大于多个音频片段的个数的正整数。

还需要说明的是，任意两个音频片段对应的声纹向量之间的相似度越高，说明任意两个音频片段对应的音频对象是同一个音频对象的可能性较高。反之，任意两个音频片段对应的

声纹向量之间的相似度越低，说明任意两个音频片段对应的音频对象是同一个音频对象的可能性较低。

示例性地，多个音频片段的个数为 5 个，则根据各个音频片段对应的声纹向量，按照上述公式（1）确定任意两个音频片段的声纹向量之间的相似度。进而根据任意两个音频片段的声纹向量之间的相似度，确定的初始相似度矩阵为 5*5 的矩阵。初始相似度矩阵如下所示：

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & a_{1,5} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & a_{2,5} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,5} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & a_{4,5} \\ a_{5,1} & a_{5,2} & a_{5,3} & a_{5,4} & a_{5,5} \end{bmatrix}。$$

在上述初始相似度矩阵中，$a_{1,1}$用于表示第 1 个音频片段对应的声纹向量和第 1 个音频片段对应的声纹向量之间的相似度，$a_{1,2}$用于表示第 1 个音频片段对应的声纹向量和第 2 个音频片段对应的声纹向量之间的相似度，初始相似度矩阵中的其他元素所代表的含义与$a_{1,1}$、$a_{1,2}$所代表的含义类似，在此不再进行一一赘述。

在步骤 203 中，根据初始相似度矩阵中各行对应的动态阈值，对初始相似度矩阵进行调整，得到参考相似度矩阵。

其中，动态阈值用于对不同相似度之间的相似度差值进行调节。动态阈值用于拉近同一音频对象的音频片段的声纹向量的相似度之间的差值，和/或，拉远不同音频对象的音频片段的声纹向量的相似度之间的差值。可选地，拉近同一音频对象的音频片段的声纹向量的相似度之间的差值是指拉近第一相似度和第二相似度之间的差值，其中，第一相似度为第一音频片段的声纹向量和第二音频片段的声纹向量之间的相似度，第二相似度为第一音频片段的声纹向量和第三音频片段的声纹向量之间的相似度，第一音频片段、第二音频片段、第三音频片段对应于同一个音频对象。拉远不同音频对象的音频片段的声纹向量的相似度之间的差值是指拉远第一相似度和第三相似度之间的差值，第三相似度为第一音频片段的声纹向量和第四音频片段的声纹向量之间的相似度，第一音频片段、第四音频片段对应于不同的音频对象。

在一种可能的实现方式中，根据初始相似度矩阵中各行对应的动态阈值，对初始相似度矩阵进行调整，得到参考相似度矩阵之前，还需确定初始相似度矩阵中各行对应的动态阈值。该过程包括：对于初始相似度矩阵中的任一行，按照第一顺序对任一行包括的多个相似度中位于预设相似度范围的相似度进行排序，得到第一排序结果；根据第一排序结果，确定位于预设相似度范围内的相似度中相邻的两个相似度之间的相似度差值，得到多个相似度差值，相似度差值的个数小于任一行包括的相似度的个数，也即，确定第一排序结果中相邻的两个相似度之间的相似度差值，得到多个相似度差值；在多个相似度差值中确定满足第一要求的相似度差值；根据满足第一要求的相似度差值，确定任一行对应的动态阈值。可选地，将满足第一要求的相似度差值对应的被减数作为任一行对应的动态阈值。

其中，第一顺序可以是从小到大的顺序，也可以是从大到小的顺序，本申请实施例对此不进行限定。预设相似度范围是基于经验进行设置，或者是根据实施环境进行调整得到的，本申请实施例对此也不进行限定。示例性地，预设相似度范围为[-1，1]。多个相似度差值中满足第一要求的相似度差值是指多个相似度差值中最大的相似度差值。

示例性地，第一顺序为从小到大的顺序，预设相似度范围为[-1，1]，初始相似度矩阵中任一行包括的相似度分别为：1、-0.3、0.7、0.5、0.9。将位于预设相似度范围内的相似度按照从小到大的顺序进行排序，得到的第一排序结果为：-0.3、0.5、0.7、0.9、1。根据第一排序结果，确定相邻的两个相似度之间的相似度差值，得到多个相似度差值，分别为：0.5-（-0.3）=0.8、0.7-0.5=0.2、0.9-0.7=0.2、1-0.9=0.1。其中，多个相似度差值中最大的相似度差值为 0.8，因此，将 0.8 对应的被减数 0.5 作为任一行对应的动态阈值。在一些实施例中，当存在多个与最大相似度差值对应的被减数时，确定任意一个被减数作为动态阈值。

可选地，根据第一排序结果，确定位于预设相似度范围内的相似度中相邻的两个相似度之间的相似度差值，得到多个相似度差值之后，还可以根据多个相似度差值确定相似度差值向量，相似度差值向量中包括多个相似度差值。示例性地，下述公式（2）为相似度差值向量。

$$gap_q = [a'_{q,2} - a'_{q,1}, a'_{q,3} - a'_{q,2}, ..., a'_{q,N} - a'_{q,N-1}] \quad 公式（2）$$

在上述公式（2）中，gap_q任一行对应的相似度差值向量，$a'_{q,1}$为按照从小到大的顺序排序后的第一排序结果中位于第一位的相似度，$a'_{q,2}$为按照从小到大的顺序排序后的第一排序结果中位于第二位的相似度，$a'_{q,3}$为按照从小到大的顺序排序后的第一排序结果中位于第三位的相似度，$a'_{q,N}$为按照从小到大的顺序排序后的第一排序结果中位于最后一位的相似度，$a'_{q,N-1}$为按照从小到大的顺序排序后的第一排序结果中位于倒数第二位的相似度。

在一种可能的实现方式中，确定出初始相似度矩阵中各行对应的动态阈值之后，根据初始相似度矩阵中各行对应的动态阈值，有下述两种实现方式对初始相似度矩阵进行调整，得到参考相似度矩阵。

实现方式一、将初始相似度矩阵第 k 行包括的相似度中，小于第 k 行对应的动态阈值的相似度调整为第一数值，并基于各行的调整结果得到参考相似度矩阵，k 为正整数。

其中，第一数值基于经验进行设置，或者根据实施环境进行调整，本申请实施例对此不进行限定。可选地，第一数值为 0。

示例性地，初始相似度矩阵为 $\begin{bmatrix} 1 & -0.3 & 0.7 & 0.5 & 0.9 \\ -0.3 & 1 & -0.5 & -0.7 & 0.6 \\ 0.7 & -0.5 & 1 & 0.8 & 0.4 \\ 0.5 & -0.7 & 0.8 & 1 & 0.2 \\ 0.9 & 0.6 & 0.4 & 0.2 & 1 \end{bmatrix}$。其中，第一行对应的动态阈值为 0.5，第二行对应的动态阈值为 0.6，第三行对应的动态阈值为 0.7，第四行对应的动态阈值为 0.2，第五行对应的动态阈值为 0.9，则将初始相似度矩阵各行包括的相似度中，小于各行对应的动态阈值的相似度调整为 0，得到的参考相似度矩阵为 $\begin{bmatrix} 1 & 0 & 0.7 & 0.5 & 0.9 \\ 0 & 1 & 0 & 0 & 0.6 \\ 0.7 & 0 & 1 & 0.8 & 0 \\ 0.5 & 0 & 0.8 & 1 & 0.2 \\ 0.9 & 0 & 0 & 0 & 1 \end{bmatrix}$。

实现方式二、将初始相似度矩阵第 k 行包括的相似度中，小于第 k 行对应的动态阈值的相似度与第二数值相乘，并基于各行的调整结果得到参考相似度矩阵。

其中，第二数值基于经验进行设置，或者根据实施环境进行调整，本申请实施例对此不进行限定。可选地，第二数值为 0.01。

示例性地，初始相似度矩阵为 $\begin{bmatrix} 1 & -0.3 & 0.7 & 0.5 & 0.9 \\ -0.3 & 1 & -0.5 & -0.7 & 0.6 \\ 0.7 & -0.5 & 1 & 0.8 & 0.4 \\ 0.5 & -0.7 & 0.8 & 1 & 0.2 \\ 0.9 & 0.6 & 0.4 & 0.2 & 1 \end{bmatrix}$。其中，第一行对应的动态阈值为 0.5，第二行对应的动态阈值为 0.6，第三行对应的动态阈值为 0.7，第四行对应的动态阈值为 0.2，第五行对应的动态阈值为 0.9，则将初始相似度矩阵各行包括的相似度中，小于各行对应的动态阈值的相似度与 0.01 相乘，得到的参考相似度矩阵为

$$\begin{bmatrix} 1 & -0.003 & 0.7 & 0.5 & 0.9 \\ -0.003 & 1 & -0.005 & -0.007 & 0.6 \\ 0.7 & -0.005 & 1 & 0.8 & 0.004 \\ 0.5 & -0.007 & 0.8 & 1 & 0.2 \\ 0.9 & 0.006 & 0.004 & 0.002 & 1 \end{bmatrix}$$。

需要说明的是，参考相似度矩阵中第一相似度和第二相似度之间的距离小于初始相似度矩阵中第一相似度和第二相似度之间的距离，参考相似度矩阵中第一相似度和第三相似度之间的距离大于初始相似度矩阵中第一相似度和第三相似度之间的距离，以达到拉近同一音频对象的音频片段的声纹向量的相似度之间的差值，拉远不同音频对象的音频片段的声纹向量之间的差值。

在步骤 204 中，根据参考相似度矩阵确定发声得到多个音频片段的音频对象的数目。

在一种可能的实现方式中，根据参考相似度矩阵确定多个音频片段中存在的音频对象的数目的过程包括：根据多个参考参数，对参考相似度矩阵进行处理，得到各个参考参数对应的相似度矩阵；根据多个参考参数和各个参考参数对应的相似度矩阵，确定多个音频片段中存在的音频对象的数目。其中，参考参数基于经验进行设置，或者根据实施环境进行调整，本申请实施例对此不进行限定。参考参数的个数本申请也不进行限定。

可选地，根据多个参考参数，对参考相似度矩阵进行处理，得到各个参考参数对应的相似度矩阵的过程是类似的，本申请实施例仅以多个参考参数中的任一个参考参数对应的相似度矩阵的确定过程为例进行说明，该过程包括下述步骤 1 至步骤 5。

步骤 1、根据任一参考参数，对参考相似度矩阵进行数值调整，得到第一相似度矩阵，数值调整用于简化参考相似度矩阵。

在一种可能的实现方式中，根据任一参考参数，有下述两种方式对参考相似度矩阵进行数值调整，得到第一相似度矩阵。

方式一、对于参考相似度矩阵各行包括的多个相似度，将任一参考参数个满足第三要求的相似度之外的相似度调整为第三数值，得到第一相似度矩阵。

其中，第三数值基于经验进行设置，或者根据实施环境进行调整，本申请实施例对此不进行限定。示例性地，第三数值为 0。满足第三要求的任一参考参数个相似度是指任一参考参数个最大的相似度，也即，将与任意参考参数对应数量的相似度调整为第三数值，且被调整为第三数值的相似度是参考相似度矩阵行中最大的相似度。

可选地，将参考相似度矩阵各行包括的多个相似度分别按照从大到小的顺序进行排序，得到各行对应的排序结果，将各行对应的排序结果中，除了前任一参考参数个相似度之外的相似度调整为第三数值，得到第一相似度矩阵。

示例性地，任一参考参数为 3，第三数值为 0，参考相似度矩阵为 $\begin{bmatrix} 1 & -0.003 & 0.7 & 0.5 & 0.9 \\ -0.003 & 1 & -0.005 & -0.007 & 0.6 \\ 0.7 & -0.005 & 1 & 0.8 & 0.004 \\ 0.5 & -0.007 & 0.8 & 1 & 0.2 \\ 0.9 & 0.006 & 0.004 & 0.002 & 1 \end{bmatrix}$。根据任一参考参数，确定第一行中满足第三要求的 3 个相似度为 1、0.9、0.7，第二行中满足第三要求的 3 个相似度为 1、0.6、-0.003，第三行中满足第三要求的 3 个相似度为 1、0.8、0.7，第四行中满足第三要求的 3 个相似度为 1、0.8、0.5，第五行中满足第三要求的 3 个相似度为 1、0.9、0.006。对参考相似度矩阵进行

调整，得到第一相似度矩阵为 $\begin{bmatrix} 1 & 0 & 0.7 & 0 & 0.9 \\ -0.003 & 1 & 0 & 0 & 0.6 \\ 0.7 & 0 & 1 & 0.8 & 0 \\ 0.5 & 0 & 0.8 & 1 & 0 \\ 0.9 & 0.006 & 0 & 0 & 1 \end{bmatrix}$。

方式二、将参考相似度矩阵包括的多个相似度中，除任一参考参数个满足第三要求的相似度之外的相似度与第四数值相乘，得到第一相似度矩阵。

其中，第四数值基于经验进行设置，或者根据实施环境进行调整，本申请实施例对此不进行限定。示例性地，第四数值为 0.01。

可选地，将参考相似度矩阵各行包括的多个相似度分别按照从大到小的顺序进行排序，得到各行对应的排序结果，将各行对应的排序结果中除前任一参考参数个相似度之外的相似度与第四数值相乘，得到第一相似度矩阵，也即，将与任意参考参数对应数量的相似度以外的相似度与第四数值相乘，且与任意参考参数对应数量的相似度是参考相似度矩阵行中最大的相似度。

示例性地，任一参考参数为 3，第四数值为 0.01，参考相似度矩阵为 $\begin{bmatrix} 1 & -0.003 & 0.7 & 0.5 & 0.9 \\ -0.003 & 1 & -0.005 & -0.007 & 0.6 \\ 0.7 & -0.005 & 1 & 0.8 & 0.004 \\ 0.5 & -0.007 & 0.8 & 1 & 0.2 \\ 0.9 & 0.006 & 0.004 & 0.002 & 1 \end{bmatrix}$。根据任一参考参数，确定第一行中满足第三要求的 3 个相似度为 1、0.9、0.7，第二行中满足第三要求的 3 个相似度为 1、0.6、-0.003，第三行中满足第三要求的 3 个相似度为 1、0.8、0.7，第四行中满足第三要求的 3 个相似度为 1、0.8、0.5，第五行中满足第三要求的 3 个相似度为 1、0.9、0.006。对参考相似度矩阵进行调整，得到第一相似度矩阵为参考相似度矩阵为 $\begin{bmatrix} 1 & -0.00003 & 0.7 & 0.005 & 0.9 \\ -0.003 & 1 & -0.00005 & -0.00007 & 0.6 \\ 0.7 & -0.00005 & 1 & 0.8 & 0.00004 \\ 0.5 & -0.00007 & 0.8 & 1 & 0.002 \\ 0.9 & 0.006 & 0.00004 & 0.00002 & 1 \end{bmatrix}$。

需要说明的是，可以选择上述任一种方式对参考相似度矩阵进行数值调整，得到第一相似度矩阵，本申请实施例对此不进行限定。

可选地，按照下述公式（3）对参考相似度矩阵进行数值调整，得到第一相似度矩阵。

$$B = Threshold(A,p) \quad \text{公式（3）}$$

在上述公式（3）中，**B** 为第一相似度矩阵，**A** 为参考相似度矩阵，p 为任一参考参数，Threshold 为数值调整函数。

步骤 2、对第一相似度矩阵进行对称化处理，得到第二相似度矩阵，所述第二相似度矩阵中位于第 i 行第 j 列的相似度与位于第 j 行第 i 列的相似度相同，i 和 j 为不大于多个音频片段的个数的正整数。

由于初始相似度矩阵是对称化矩阵，经过动态阈值处理和数值调整后的第一相似度矩阵可能为非对称化矩阵，故对第一相似度矩阵进行对称化处理。由于第 i 个音频片段的声纹向量与第 j 个音频片段的声纹向量之间的相似度、第 j 个音频片段的声纹向量与第 i 个音频片段的声纹向量之间的相似度是相同的，也即是位于第 i 行第 j 列的相似度与位于第 j 行第 i 列的相似度是相同的，因此，需要对第一相似度矩阵进行对称化处理，以使位于第 i 行第 j 列的相似度与位于第 j 行第 i 列的相似度是相同的。可选地，至少存在下述方式对第一相似度矩阵进

行对称化处理，得到第二相似度矩阵。

方式 1、确定第一相似度矩阵对应的转置矩阵；将第一相似度矩阵和第一相似度矩阵对应的转置矩阵中位于相同位置的相似度相加，得到待调整相似度矩阵；对待调整相似度矩阵包括的多个相似度进行取半操作，得到第二相似度矩阵。

可选地，按照下述公式（4）对第一相似度矩阵进行对称化处理，得到第二相似度矩阵。

$$C=\frac{1}{2}(B+B^T) \quad \text{公式（4）}$$

在上述公式（4）中，**C** 为第二相似度矩阵，**B** 为第一相似度矩阵，B^T 为第一相似度矩阵对应的转置矩阵。

示例性地，第一相似度矩阵为 $\begin{bmatrix} 1 & 0 & 0.7 & 0 & 0.9 \\ -0.003 & 1 & 0 & 0 & 0.6 \\ 0.7 & 0 & 1 & 0.8 & 0 \\ 0.5 & 0 & 0.8 & 1 & 0 \\ 0.9 & 0.006 & 0 & 0 & 1 \end{bmatrix}$，第一相似度矩阵对应的转置矩阵为 $\begin{bmatrix} 1 & -0.003 & 0.7 & 0.5 & 0.9 \\ 0 & 1 & 0 & 0 & 0.006 \\ 0.7 & 0 & 1 & 0.8 & 0 \\ 0 & 0 & 0.8 & 1 & 0 \\ 0.9 & 0.6 & 0 & 0 & 1 \end{bmatrix}$。将第一相似度矩阵和第一相似度矩阵对应的转置矩阵中位于相同位置的相似度相加，得到的待调整相似度矩阵为 $\begin{bmatrix} 2 & -0.003 & 1.4 & 0.5 & 1.8 \\ -0.003 & 2 & 0 & 0 & 0.606 \\ 1.4 & 0 & 2 & 1.6 & 0 \\ 0.5 & 0 & 1.6 & 2 & 0 \\ 1.8 & 0.606 & 0 & 0 & 2 \end{bmatrix}$，对待调整相似度矩阵中包括的多个相似度进行取半操作，得到的第二相似度矩阵为 $\begin{bmatrix} 1 & -0.0015 & 0.7 & 0.25 & 0.9 \\ -0.0015 & 1 & 0 & 0 & 0.303 \\ 0.7 & 0 & 1 & 0.8 & 0 \\ 0.25 & 0 & 0.8 & 1 & 0 \\ 0.9 & 0.303 & 0 & 0 & 1 \end{bmatrix}$。

方式 1 为根据第一相似度矩阵和第一相似度矩阵对应的转置矩阵，通过取平均值的方式，来确定第二相似度矩阵的过程。

方式 2、确定第一相似度矩阵中位于第 i 行第 j 列的相似度、与第一相似度矩阵中位于第 j 行第 i 列的相似度中最大的相似度，将最大的相似度作为第二相似度矩阵中位于第 i 行第 j 列和第 j 行第 i 列的相似度，得到第二相似度矩阵。

可选地，按照下述公式（5）对第一相似度矩阵进行对称化处理，得到第二相似度矩阵。

$$a'_{ij}=\max(a_{ij},a_{ji}) \quad \text{公式（5）}$$

在上述公式（5）中，$a'_{i,j}$ 为第二相似度矩阵中位于第 i 行、第 j 列的相似度，a_{ij} 为第一相似度矩阵中位于第 i 行、第 j 列的相似度，a_{ji} 为第一相似度矩阵中位于第 j 行、第 i 列的相似度。

示例性地，第一相似度矩阵为 $\begin{bmatrix} 1 & 0 & 0.7 & 0 & 0.9 \\ -0.003 & 1 & 0 & 0 & 0.6 \\ 0.7 & 0 & 1 & 0.8 & 0 \\ 0.5 & 0 & 0.8 & 1 & 0 \\ 0.9 & 0.006 & 0 & 0 & 1 \end{bmatrix}$，则第二相似度矩阵为 $\begin{bmatrix} 1 & 0 & 0.7 & 0.5 & 0.9 \\ 0 & 1 & 0 & 0 & 0.6 \\ 0.7 & 0 & 1 & 0.8 & 0 \\ 0.5 & 0 & 0.8 & 1 & 0 \\ 0.9 & 0.6 & 0 & 0 & 1 \end{bmatrix}$。

方式 2 为根据第一相似度矩阵，通过取最大值的方式，确定第二相似度矩阵的过程。

需要说明的是，可以选择上述任一种方式对第一相似度矩阵进行对称化处理，得到第二相似度矩阵，本申请实施例对此不进行限定。

步骤 3、对第二相似度矩阵进行行列扩散，得到第三相似度矩阵，第三相似度矩阵用于生成多个音频对象之间的边界。

在一种可能的实现方式中，对第二相似度矩阵进行行列扩散，得到第三相似度矩阵的过程包括：确定第二相似度矩阵对应的转置矩阵，根据第二相似度矩阵和第二相似度矩阵对应的转置矩阵，确定第三相似度矩阵，第三相似度矩阵中位于第 m 行第 n 列的相似度基于第二相似度矩阵中位于第 m 行的相似度和第二相似度矩阵对应的转置矩阵中位于第 n 列的相似度确定，m、n 为不大于多个音频片段的个数的正整数。

可选地，对于第三相似度矩阵中位于第 m 行第 n 列的相似度，将第二相似度矩阵中位于第 m 行的相似度和第二相似度矩阵对应的转置矩阵中位于第 n 列的相似度对应相乘再相加的结果作为第三相似度矩阵中位于第 m 行第 n 列的相似度。

示例性地，第二相似度矩阵中位于第 m 行的相似度分别为 1、0、0.7、0.5、0.9，第二相似度矩阵对应的转置矩阵中位于第 n 列的相似度分别为 1、0、0.7、0.5、0.9，则第三相似度矩阵中位于第 m 行第 n 列的相似度为 1*1+0*0+0.7*0.7+0.5*0.5+0.9*0.9=2.55。

需要说明的是，第三相似度矩阵中其他位置的相似度的确定过程与上述第 m 行第 n 列的相似度的确定过程类似，在此不再进行赘述。

可选地，按照下述公式（6）对第二相似度矩阵进行行列扩散，得到第三相似度矩阵。

$$D = CC^T \quad \text{公式（6）}$$

在上述公式（6）中，**D** 为第三相似度矩阵，**C** 为第二相似度矩阵，C^T 为第二相似度矩阵对应的转置矩阵。

示例性地，第二相似度矩阵为 $\begin{bmatrix} 1 & 0 & 0.7 & 0.5 & 0.9 \\ 0 & 1 & 0 & 0 & 0.6 \\ 0.7 & 0 & 1 & 0.8 & 0 \\ 0.5 & 0 & 0.8 & 1 & 0 \\ 0.9 & 0.6 & 0 & 0 & 1 \end{bmatrix}$，第二相似度矩阵对应的转置矩阵为 $\begin{bmatrix} 1 & 0 & 0.7 & 0.5 & 0.9 \\ 0 & 1 & 0 & 0 & 0.6 \\ 0.7 & 0 & 1 & 0.8 & 0 \\ 0.5 & 0 & 0.8 & 1 & 0 \\ 0.9 & 0.6 & 0 & 0 & 1 \end{bmatrix}$，则第三相似度矩阵为

$$\begin{bmatrix} 2.55 & 0.54 & 1.8 & 0.56 & 1.8 \\ 0.54 & 1.36 & 0 & 0 & 1.2 \\ 1.8 & 0 & 2.13 & 1.95 & 0.63 \\ 1.56 & 0 & 1.95 & 1.89 & 0.45 \\ 1.8 & 1.2 & 0.63 & 0.45 & 2.17 \end{bmatrix}。$$

步骤 4、对第三相似度矩阵进行比例调整，得到第四相似度矩阵，比例调整用于将第三相似度矩阵中各行包括的相似度调整在同一范围内。

在一种可能的实现方式中，对第三相似度矩阵进行比例调整，得到第四相似度矩阵的过程包括：根据第三相似度矩阵中各行包括的多个相似度，确定各行对应的最大相似度；将第三相似度矩阵中各行包括的多个相似度分别与各行对应的最大相似度相除，得到第四相似度矩阵。

可选地，按照下述公式（7）对第三相似度矩阵进行比例调整，得到第四相似度矩阵。

$$a''_{ij} = a_{ij} / a_{ik} \quad \text{公式（7）}$$

在上述公式（7）中，a''_{ij} 为第四相似度矩阵中位于第 i 行第 j 列的相似度，a_{ij} 为第三相似度矩阵中位于第 i 行第 j 列的相似度，a_{ik} 为第三相似度矩阵中第 i 行对应的最大相似度，k 为第三相似度矩阵中第 i 行对应的最大相似度所在的列。

示例性地，第三相似度矩阵为 $\begin{bmatrix} 2.55 & 0.54 & 1.8 & 0.56 & 1.8 \\ 0.54 & 1.36 & 0 & 0 & 1.2 \\ 1.8 & 0 & 2.13 & 1.95 & 0.63 \\ 1.56 & 0 & 1.95 & 1.89 & 0.45 \\ 1.8 & 1.2 & 0.63 & 0.45 & 2.17 \end{bmatrix}$，其中，第一行对应的最大相似度为 2.55，第二行对应的最大值为 1.36，第三行对应的最大值为 2.13，第四行对应的最大值为 1.95，第五行对应的最大值为 2.17。根据各行对应的最大相似度，对第三相似度矩阵进行比例调整，得到的第四相似度矩阵为 $\begin{bmatrix} 1 & 0.212 & 0.76 & 0.612 & 0.706 \\ 0.397 & 1 & 0 & 0 & 0.882 \\ 0.845 & 0 & 1 & 0.915 & 0.296 \\ 0.8 & 0 & 1 & 0.969 & 0.231 \\ 0.829 & 0.553 & 0.29 & 0.207 & 1 \end{bmatrix}$。

步骤 5、对第四相似度矩阵进行对称化处理，得到任一参考参数对应的相似度矩阵。

在一种可能的实现方式中，对第四相似度矩阵进行对称化处理，得到任一参考参数对应的相似度矩阵的过程与上述对第一相似度矩阵进行对称化处理，得到第二相似度矩阵的过程是类似的，在此不再进行赘述。

需要说明的是，根据上述步骤 1 至步骤 5 的过程，分别确定出各个参考参数对应的相似度矩阵。

图 3 是本申请实施例提供的一种相似度矩阵的确定过程的示意图。图 3 中的（1）为初始相似度矩阵，图 3 中的（2）为参考相似度矩阵，图 3 中的（3）为第一相似度矩阵，图 3 中的（4）为第三相似度矩阵，图 3 中的（5）为第四相似度矩阵，图 3 中的（6）为参考参数对应的相似度矩阵。图 3 中的（1）中的横轴为音频片段的个数，纵轴为音频片段的个数，图 3 中区域亮度越高表示两个音频片段的声纹向量之间的相似度越高。

在一种可能的实现方式中，根据多个参考参数和各个参考参数对应的相似度矩阵，确定多个音频片段中存在的音频对象的数目的过程包括：根据多个参考参数和各个参考参数对应的相似度矩阵，确定各个参考参数对应的比例值，比例值用于指示参考参数对应的相似度矩阵中保留的相似度的数量；根据各个参考参数对应的比例值，确定多个音频片段中存在的音

频对象的数目。比例值越小，说明参考参数对应的相似度矩阵中保留的相似度的数量越少，后续确定的音频对象的数目的准确性越高；反之，比例值越大，说明参考参数对应的相似度矩阵中保留的相似度的数量越多，后续确定的音频对象的数目的准确性越低。

可选地，根据多个参考参数和各个参考参数对应的相似度矩阵，确定各个参考参数对应的比例值的过程包括：对于多个参考参数中的任一参考参数，对任一参考参数对应的相似度矩阵进行拉普拉斯变换，得到任一参考参数对应的拉普拉斯矩阵；对拉普拉斯矩阵进行奇异值分解，得到多个参考特征值；在多个参考特征值中确定第二特征值和第一数量个第一特征值，第二特征值为多个参考特征值中的最大值，第一特征值为按照第二顺序对多个参考特征值进行排序后满足第二要求的参考特征值。确定第一数量个第一特征值中相邻的两个第一特征值之间的差值，得到多个特征值差值；根据第二特征值，对第一特征值差值进行归一化处理，得到归一化之后的特征值差值，第一特征值差值为多个特征值差值中最大的特征值差值；根据归一化之后的特征值差值和任一参考参数，确定任一参考参数对应的比例值。其中，第一数量基于经验进行设置，或者根据实施环境进行调整，本申请实施例对此不进行限定。例如，第一数量为 3。第二顺序可以是从小到大的顺序，也可以是从大到小的顺序，本申请实施例对此不进行限定。当第二顺序为从小到大的顺序时，则第一特征值为按照从小到大的顺序对多个参考特征值进行排序后，前第一数量个参考特征值。当第二顺序为从大到小的顺序时，则第一特征值为按照从大到小的顺序对多个参考特征值进行排序后，后第一数量个参考特征值。

可选地，确定第一数量个第一特征值中相邻的两个第一特征值之间的差值，得到多个特征值差值之后，还可以根据多个特征值差值确定任一参考参数对应的特征值差值向量，特征值差值向量中包括多个特征值差值。示例性地，下述公式（8）为任一参考参数对应的特征值差值向量。

$$e_p = [\lambda_{p,2} - \lambda_{p,1}, \lambda_{p,3} - \lambda_{p,2}, \ldots, \lambda_{p,Y} - \lambda_{p,Y-1}] \quad \text{公式（8）}$$

在上述公式（8）中，e_p 为任一参考参数对应的特征值差值向量，$\lambda_{p,1}$ 为按照从小到大的顺序对多个参考特征值进行排序后位于第一位的参考特征值，$\lambda_{p,2}$ 为按照从小到大的顺序对多个参考特征值进行排序后位于第二位的参考特征值，$\lambda_{p,3}$ 为按照从小到大的顺序对多个参考特征值进行排序后位于第三位的参考特征值，$\lambda_{p,Y}$ 为按照从小到大的顺序对多个参考特征值进行排序后位于第 Y 位的参考特征值，$\lambda_{p,Y-1}$ 为按照从小到大的顺序对多个参考特征值进行排序后位于第 Y-1 位的相似度。Y 为第一数量。

可选地，根据第二特征值，按照下述公式（9）对特征值差值进行归一化处理，得到归一化之后的特征值差值。

$$g_P = \frac{\max(e_p)}{\lambda_{\max} + \varepsilon} \quad \text{公式（9）}$$

在上述公式（9）中，g_P 为归一化之后的特征值差值，$\max(e_p)$ 为第一特征值差值，$\lambda_{\max}$ 为第二特征值，ε 为归一化参数，ε 的取值为 $1*10^{-10}$。

根据归一化之后的特征值差值和任一参考参数，按照下述公式（10）确定任一参考参数对应的比例值。

$$r(p) = \frac{p}{g_p} \quad \text{公式（10）}$$

在上述公式（10）中，$r(p)$ 为任一参考参数对应的比例值，p 为任一参考参数，g_p 为归一化之后的特征值差值。

在一种可能的实现方式中，根据各个参考参数对应的比例值，确定多个音频片段中存在的音频对象的数目的过程包括：根据各个参考参数对应的比例值，在多个参考参数中确定第一参数，第一参数为多个参考参数中对应的比例值最小的参考参数；确定第一参数对应的多

个特征值差值，调用第一函数对第一参数对应的多个特征值差值进行处理，得到多个音频片段中存在的音频对象的数目。可选地，将各个参考参数对应的比例值中，最小的比例值对应的参考参数作为第一参数。

需要说明的是，确定第一参数对应的多个特征值差值的过程为：确定第一参数对应的相似度矩阵，对第一参数对应的相似度矩阵进行拉普拉斯变换，得到第一参数对应的拉普拉斯矩阵，对第一参数对应的拉普拉斯矩阵进行奇异值分解，得到多个参考特征值，将多个参考特征值中最小的第一数量个参考特征值进行排序，将排序中相邻的两个参考特征值之间的差值作为第一参数对应的多个特征值差值。

调用第一函数对第一参数对应的多个特征值差值进行处理，得到多个音频片段中存在的音频对象的数目的过程包括：将第一参数对应的多个特征值差值组成第一参数对应的特征值差值向量，调用第一函数对第一参数对应的特征值差值向量进行处理，得到多个音频片段中存在的音频对象的数目。

可选地，按照下述公式（11）对第一参数对应的多个特征值差值进行处理，得到多个音频片段中存在的音频对象的数目。

$$M = arg\max(e_{\hat{p}}) \quad \text{公式（11）}$$

在上述公式（11）中，M 为多个音频片段中存在的音频对象的数目，$arg\max(\ )$为第一函数，$e_{\hat{p}}$为第一参数对应的特征值差值向量，第一参数对应的特征值差值向量是由第一参数对应的多个特征值差值组成的向量。

示例性地，第一参数对应的相似度矩阵为矩阵 Q，对矩阵 Q 进行拉普拉斯变换，得到第一参数对应的拉普拉斯矩阵 P，对矩阵 P 进行奇异值分解，得到多个参考特征值（分别为 a、b、c、d、e、f），将多个参考特征值按照从小到大的顺序进行排序，得到排序结果（b、c、a、e、f、d），第一数量为 3，将排序结果中最小的 3 个参考值中相邻的两个参考特征值之间的差值作为第一参数对应的多个特征值差值，多个特征值差值分别为 c-b、a-c，因此，将 c-b、a-c 组成的向量作为特征值差值向量，也即是特征值差值向量为[c-b，a-c]。

在步骤 205 中，根据音频对象的数目，对多个音频片段进行聚类，得到各个音频对象发声得到的音频片段。

在一种可能的实现方式中，基于上述步骤 204 确定出多个音频片段中存在的音频对象的数目之后，根据音频对象的数目，对多个音频片段进行聚类，得到各个音频对象对应的音频片段的过程包括：对第一参数对应的相似度矩阵进行奇异值分解，得到多个分解特征值；在多个分解特征值中确定音频对象的数目个分解特征值；确定音频对象的数目个分解特征值分别对应的特征向量；根据音频对象的数目个分解特征值分别对应的特征向量，生成分解矩阵，分解矩阵的行数为音频对象的数目，列数为音频片段的个数；根据分解矩阵，确定多个音频片段分别对应的特征向量，特征向量用于指示对应的音频片段；根据音频对象的数目和多个音频片段分别对应的特征向量，对多个音频片段进行聚类，得到各个音频对象对应的音频片段。

可选地，在多个分解特征值中确定音频对象的数目个分解特征值时，确定的分解特征值是最小的音频对象的数目个分解特征值。

示例性地，音频对象的数目为 3，则在多个分解特征值中确定最小的 3 个分解特征值。确定这三个分解特征值分别对应的特征向量，分解特征值对应的特征向量为 1*5 的特征向量，将 3 个 1*5 的特征向量组成 3*5 的分解矩阵。将分解矩阵中第一列作为第一个音频片段对应的特征向量，第二列作为第二个音频片段对应的特征向量，第三列作为第三个音频片段对应的特征向量，第四列作为第四个音频片段对应的特征向量的，第五列作为第五个音频片段对应的特征向量。

示例性地，三个分解特征值分别对应的特征向量为$[x_1,x_2,x_3,x_4,x_5]$、$[y_1,y_2,y_3,y_4,y_5]$、$[z_1,z_2,z_3,z_4,z_5]$，则根据三个分解特征值分别对应的特征向量，组成的分解矩阵为

$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ y_1 & y_2 & y_3 & y_4 & y_5 \\ z_1 & z_2 & z_3 & z_4 & z_5 \end{bmatrix}$。因此，将$[x_1,y_1,z_1]$作为第一个音频片段对应的特征向量，将$[x_2,y_2,z_2]$作为第二个音频片段对应的特征向量，将$[x_3,y_3,z_3]$作为第三个音频片段对应的特征向量，将$[x_4,y_4,z_4]$作为第四个音频片段对应的特征向量，将$[x_5,y_5,z_5]$作为第五个音频片段对应的特征向量。

可选地，根据音频对象的数目和多个音频片段分别对应的特征向量，通过 K-means（K-均值）聚类算法对多个音频片段进行聚类，得到各个音频对象对应的音频片段，其中，K 的取值为音频对象的数目。当然，还可以使用其他的聚类算法对多个音频片段进行聚类，本申请实施例对此不进行限定。

示例性地，待处理的音频片段有 5 个，分别为音频片段 1、音频片段 2、音频片段 3、音频片段 4 和音频片段 5。根据上述步骤 201 至步骤 204，确定出待处理的音频片段中存在的音频对象的数目为 3。根据上述步骤 205，确定出音频对象 1 对应的音频片段为音频片段 1、音频片段 3，音频对象 2 对应的音频片段为音频片段 5，音频对象 3 对应的音频片段为音频片段 2 和音频片段 4。

在一种可能的实现方式中，在上述步骤 204 中确定出多个音频片段中存在的音频对象的数目之后，还可以根据音频对象的数目和多个音频片段分别对应的声纹向量，对多个音频片段进行聚类，得到各个音频对象对应的音频片段。

其中，根据音频对象的数目和多个音频片段分别对应的声纹向量，对多个音频片段进行聚类，得到各个音频对象对应的音频片段的过程与上述根据音频对象的数目和多个音频片段分别对应的特征向量，对多个音频片段进行聚类，得到各个音频对象对应的音频片段的过程类似，在此不再进行赘述。

本申请实施例提供的音频处理方法可应用于游戏领域中，确定同一个游戏账号（或同一个智能设备）由几个用户使用。可选地，采集用户在使用该游戏账号（或该智能设备）时的音频片段，调用本申请实施例提供的音频处理方法，确定各个音频片段对应的声纹向量，根据各个音频片段对应的声纹向量，确定音频片段中存在的音频对象的数目，以及各个音频对象对应的音频片段，以获知该游戏账号（或该智能设备）供几个用户使用。

上述方法根据初始相似度矩阵中各行对应的动态阈值，对初始相似度矩阵进行调整，进而得到参考相似度矩阵，通过动态阈值调整过程，能够拉近同一音频对象的音频片段的声纹向量的相似度，拉远不同音频对象的音频片段的声纹向量的相似度，使得根据参考相似度矩阵，确定的音频对象的数目更加准确；再根据准确率较高的音频对象的数目，对多个音频片段进行聚类，得到各个音频对象对应的音频片段，使得确定的各个音频对象对应的音频片段的准确性较高，音频对象聚类的准确率较高，进而能够提高音频片段的音频处理效果。

本实施例提供的方法，通过对初始相似度矩阵中的人一行进行相似度排序，从而根据相似度排序中各相邻相似度之间的差值确定动态阈值，从而以动态阈值对初始相似度矩阵进行调整，提高了拉近属于相同音频对象的声纹相似度以及拉远不同音频对象的声纹相似度的准确率，提高了确定音频对象数量的准确率。

本实施例提供的方法，基于动态阈值对初始相似度矩阵中的相似度进行第一数值的设置，提高了初始相似度矩阵的调整效率。

本实施例提供的方法，基于动态阈值对初始相似度矩阵中的相似度与第二数值进行运算设置，提高了调整后的参考相似度矩阵的设置灵活性和准确率。

本实施例提供的方法，生成多个参考参数，并基于多个参考参数对参考相似度矩阵进行处理，得到各个参考参数对应的相似度矩阵，从而提高了相似度矩阵的准确率，并提高了确定音频对象数目的准确率。

本实施例提供的方法，在确定参考参数对应的相似度矩阵时，对参考相似度矩阵进行调

整后，由于调整后的第一相似度矩阵存在不对称的可能，故对第一相似度矩阵进行对称化处理，并再进行行列扩散和比例调整处理，提高了确定相似度矩阵的准确率。

本实施例提供的方法，在确定第一相似度矩阵时，根据第三要求确定任一参考参数个相似度，并将其他相似度进行数值调整，并通过第一相似度矩阵和对应的转置矩阵确定第二相似度矩阵，提高了参考参数对应的相似度矩阵的准确率。

图 4 是本申请实施例提供的另一种音频处理方法的流程图，如图 4 所示，该方法包括下述步骤 401 至步骤 415。

401、获取多个音频片段。

在一种可能的实现方式中，该过程已在上述步骤 201 中进行描述，在此不再进行赘述。

402、对各个音频片段进行信号预处理，得到各个音频片段的特征。

在一种可能的实现方式中，该过程已在上述步骤 201 中进行描述，在此不再进行赘述。其中，对音频片段进行信号预处理的方式包括分割、降噪、采样、量化等处理方式中的至少一种。

403、调用声纹提取模型对各个音频片段的特征进行处理，得到各个音频片段对应的声纹向量。

在一种可能的实现方式中，该过程已在上述步骤 201 中进行描述，在此不再进行赘述。示例性地，声纹提取模型可以为 CLDNN 模型，也可以为基于 TDNN 的 X-vector，还可以是 ecapa-tdnn。

404、根据各个音频片段对应的声纹向量，确定初始相似度矩阵。

在一种可能的实现方式中，该过程已在上述步骤 202 中进行描述，在此不再进行赘述。

其中，确定任意两个音频片段对应的声纹向量之间的相似度，并构建初始相似度矩阵，其中，初始相似度矩阵的纵向和横向的数量与音频片段的数量对应，从而构建各个音频片段的声纹向量之间的相似度的对应矩阵作为初始相似度矩阵。

405、根据初始相似度矩阵中各行对应的动态阈值，对初始相似度矩阵进行调整，得到参考相似度矩阵。

在一种可能的实现方式中，该过程已在上述步骤 203 中进行描述，在此不再进行赘述。其中，动态阈值用于拉近相同发声对象对应的音频片段的声纹特征之间的相似度，或者，动态阈值用于拉远不同发声对象对应的音频片段的声纹特征之间的相似度。

406、根据多个参考参数，对参考相似度矩阵进行数值调整，得到各个参考参数对应的第一相似度矩阵。

在一种可能的实现方式中，该过程已在上述步骤 204 中进行描述，在此不再进行赘述。可选地，对于参考相似度矩阵各行包括的多个相似度，将任一参考参数个满足第三要求的相似度之外的相似度调整为第三数值，得到第一相似度矩阵；或者，将参考相似度矩阵包括的多个相似度中，除任一参考参数个满足第三要求的相似度之外的相似度与第四数值相乘，得到第一相似度矩阵。

407、对各个参考参数对应的第一相似度矩阵进行对称化处理，得到各个参考参数对应的第二相似度矩阵。

在一种可能的实现方式中，该过程已在上述步骤 204 中进行描述，在此不再进行赘述。由于参考相似度矩阵被调整为第一相似度矩阵时，第一相似度矩阵存在不对称情况，故对第一相似度矩阵进行对称化处理，得到第二相似度矩阵。

408、对各个参考参数对应的第二相似度矩阵进行行列扩散，得到各个参考参数对应的第三相似度矩阵。

在一种可能的实现方式中，该过程已在上述步骤 204 中进行描述，在此不再进行赘述。其中，第二相似度矩阵通过转置矩阵进行行列扩散，得到第三转置矩阵。

409、对各个参考参数对应的第三相似度矩阵进行比例调整，得到各个参考参数对应的第

四相似度矩阵。

在一种可能的实现方式中，该过程已在上述步骤 204 中进行描述，在此不再进行赘述。可选地，根据第三相似度矩阵中各行包括的多个相似度，确定各行对应的最大相似度；将第三相似度矩阵中各行包括的多个相似度分别与各行对应的最大相似度相除，得到第四相似度矩阵。

410、对各个参考参数对应的第四相似度矩阵进行对称化处理，得到各个参考参数对应的相似度矩阵。

在一种可能的实现方式中，该过程已在上述步骤 204 中进行描述，在此不再进行赘述。由于第三相似度矩阵进行比例调整后，第四相似度矩阵存在不对称情况，故对第四相似度矩阵进行对称化处理，得到参考参数对应的相似度矩阵。

411、根据多个参考参数和各个参考参数对应的相似度矩阵，确定各个参考参数对应的比例值。

在一种可能的实现方式中，该过程已在上述步骤 204 中进行描述，在此不再进行赘述。比例值用于指示参考参数对应的相似度矩阵中保留的相似度的数量。比例值越小，说明参考参数对应的相似度矩阵中保留的相似度的数量越少，后续确定的音频对象的数目的准确性越高；反之，比例值越大，说明参考参数对应的相似度矩阵中保留的相似度的数量越多，后续确定的音频对象的数目的准确性越低。

412、根据各个参考参数对应的比例值，在多个参考参数中确定第一参数。

在一种可能的实现方式中，该过程已在上述步骤 204 中进行描述，在此不再进行赘述。

413、根据第一参数，确定多个音频片段中存在的音频对象的数目。

在一种可能的实现方式中，该过程已在上述步骤 204 中进行描述，在此不再进行赘述。

414、根据第一参数，确定各个音频片段对应的特征向量。

在一种可能的实现方式中，该过程已在上述步骤 205 中进行描述，在此不再进行赘述。

415、根据音频对象的数目和各个音频片段对应的特征向量，对多个音频片段进行聚类，得到各个音频对象对应的音频片段。

在一种可能的实现方式中，该过程已在上述步骤 205 中进行描述，在此不再进行赘述。

图 5 所示为本申请实施例提供的一种音频处理装置的结构示意图，如图 5 所示，该装置包括：

确定模块 501，用于确定多个音频片段分别对应的声纹向量，声纹向量用于指示音频片段对应的声纹特征；

确定模块 501，还用于根据各个音频片段对应的声纹向量，确定初始相似度矩阵，初始相似度矩阵中包括任意两个音频片段对应的声纹向量之间的相似度；

调整模块 502，用于根据初始相似度矩阵中各行对应的动态阈值，对初始相似度矩阵进行调整，得到参考相似度矩阵，所述动态阈值用于对不同相似度之间的相似度差值进行调节；

确定模块 501，还用于根据所述参考相似度矩阵确定发声得到所述多个音频片段的音频对象的数目；

聚类模块 503，用于根据所述音频对象的数目，对所述多个音频片段进行聚类，得到各个音频对象发声得到的音频片段。

在一种可能的实现方式中，确定模块 501，还用于对于所述初始相似度矩阵中的任一行，按照第一顺序对所述任一行中位于预设相似度范围内的相似度进行排序，得到第一排序结果；确定所述第一排序结果中相邻的两个相似度之间的相似度差值，得到多个相似度差值；在多个相似度差值中确定满足第一要求的相似度差值；根据满足第一要求的相似度差值，确定任一行对应的动态阈值。

在一种可能的实现方式中，调整模块 502，用于将所述初始相似度矩阵第 k 行包括的相似度中，小于第 k 行对应的动态阈值的相似度调整为第一数值，并基于各行的调整结果得到

所述参考相似度矩阵，k 为正整数；或者，将所述初始相似度矩阵第 k 行包括的相似度中，小于所述第 k 行对应的动态阈值的相似度与第二数值相乘，并基于各行的调整结果得到所述参考相似度矩阵。

在一种可能的实现方式中，确定模块 501，用于根据多个参考参数，对参考相似度矩阵进行处理，得到各个参考参数对应的相似度矩阵；根据多个参考参数和各个参考参数对应的相似度矩阵，确定多个音频片段中存在的音频对象的数目。

在一种可能的实现方式中，确定模块 501，用于对于多个参考参数中的任一参考参数，根据任一参考参数，对参考相似度矩阵进行数值调整，得到第一相似度矩阵，数值调整用于简化参考相似度矩阵；对第一相似度矩阵进行对称化处理，得到第二相似度矩阵，第二相似度矩阵中位于第 i 行第 j 列的相似度与位于第 j 行第 i 列的相似度相同，i 和 j 为不大于多个音频片段的个数的正整数；对第二相似度矩阵进行行列扩散，得到第三相似度矩阵，第三相似度矩阵用于生成多个音频对象之间的边界；对第三相似度矩阵进行比例调整，得到第四相似度矩阵，比例调整用于将第三相似度矩阵中各行包括的相似度调整在同一个范围内；对第四相似度矩阵进行对称化处理，得到任一参考参数对应的相似度矩阵。

在一种可能的实现方式中，确定模块 501，用于对于参考相似度矩阵各行包括的多个相似度，将任一参考参数个满足第三要求的相似度之外的相似度调整为第三数值，得到第一相似度矩阵；或者，将参考相似度矩阵包括的多个相似度中，除任一参考参数个满足第三要求的相似度之外的相似度与第四数值相乘，得到第一相似度矩阵。

在一种可能的实现方式中，确定模块 501，用于确定第一相似度矩阵对应的转置矩阵；将第一相似度矩阵和第一相似度矩阵对应的转置矩阵中位于相同位置的相似度相加，得到待调整相似度矩阵；对待调整相似度矩阵包括的多个相似度进行取半操作，得到第二相似度矩阵。

在一种可能的实现方式中，确定模块 501，用于确定第一相似度矩阵中位于第 i 行第 j 列的相似度，与第一相似度矩阵中位于第 j 行第 i 列的相似度中最大的相似度，将最大的相似度作为第二相似度矩阵中位于第 i 行第 j 列和第 j 行第 i 列的相似度，得到第二相似度矩阵。

在一种可能的实现方式中，确定模块 501，用于确定第二相似度矩阵对应的转置矩阵；根据第二相似度矩阵和第二相似度矩阵对应的转置矩阵，确定第三相似度矩阵，第三相似度矩阵中位于第 m 行第 n 列的相似度基于第二相似度矩阵中位于第 m 行的相似度和第二相似度矩阵对应的转置矩阵中位于第 n 列的相似度确定，m、n 为不大于多个音频片段的个数的正整数。

在一种可能的实现方式中，确定模块 501，用于根据第三相似度矩阵中各行包括的多个相似度，确定各行对应的最大相似度；将第三相似度矩阵中各行包括的多个相似度分别与各行对应的最大相似度相除，得到第四相似度矩阵。

在一种可能的实现方式中，确定模块 501，用于根据多个参考参数和各个参考参数对应的相似度矩阵，确定各个参考参数对应的比例值，比例值用于指示参考参数对应的相似度矩阵中保留的相似度的数量；根据各个参考参数对应的比例值，确定多个音频片段中存在的音频对象的数目。

在一种可能的实现方式中，确定模块 501，用于对于多个参考参数中的任一参考参数，对任一参考参数对应的相似度矩阵进行拉普拉斯变换，得到任一参考参数对应的拉普拉斯矩阵；对拉普拉斯矩阵进行奇异值分解，得到多个参考特征值；在多个参考特征值中确定第二特征值和第一数量个第一特征值，第二特征值为多个参考特征值中的最大值，第一特征值为按照第二顺序对多个参考特征值进行排序后满足第二要求的参考特征值；确定第一数量个第一特征值中相邻的两个第一特征值之间的差值，得到多个特征值差值；根据第二特征值，对第一特征值差值进行归一化处理，得到归一化之后的特征值差值，第一特征值差值为多个特征值差值中最大的特征值差值；根据归一化之后的特征值差值和任一参考参数，确定任一参考参数对应的比例值。

在一种可能的实现方式中，确定模块 501，用于根据各个参考参数对应的比例值，在多个参考参数中确定第一参数，第一参数为多个参考参数中对应的比例值最小的参考参数；确定第一参数对应的多个特征值差值；调用第一函数对第一参数对应的多个特征值差值进行处理，得到多个音频片段中存在的音频对象的数目。

在一种可能的实现方式中，聚类模块 503，用于对所述第一参数对应的相似度矩阵进行奇异值分解，得到多个分解特征值；在所述多个分解特征值中确定与所述音频对象的数目对应分解特征值；确定所述音频对象的数目个分解特征值分别对应的特征向量，并生成分解矩阵，所述分解矩阵的行数为所述音频对象的数目，列数为所述音频片段的数目；根据所述分解矩阵，确定所述多个音频片段分别对应的特征向量，所述特征向量用于指示对应的音频片段；根据所述音频对象的数目和所述多个音频片段分别对应的特征向量，对所述多个音频片段进行聚类，各个音频对象发声得到的音频片段。

上述装置根据初始相似度矩阵中各行对应的动态阈值，对初始相似度矩阵进行调整，进而得到参考相似度矩阵，通过动态阈值调整过程，能够拉近同一音频对象的音频片段的声纹向量的相似度，拉远不同音频对象的音频片段的声纹向量的相似度，使得根据参考相似度矩阵，确定的音频对象的数目更加准确；再根据准确率较高的音频对象的数目，对多个音频片段进行聚类，得到各个音频对象对应的音频片段，使得确定的各个音频对象对应的音频片段的准确性较高，音频对象聚类的准确率较高，进而能够提高音频片段的音频处理效果。

图 6 示出了本申请一个示例性实施例提供的终端设备 600 的结构框图。该终端设备 600 可以是便携式移动终端，比如：智能手机、平板电脑、MP3 播放器（Moving Picture Experts Group Audio Layer III，动态影像专家压缩标准音频层面 3）、MP4（Moving Picture Experts Group Audio Layer IV，动态影像专家压缩标准音频层面 4）播放器、笔记本电脑或台式电脑。终端设备 600 还可能被称为用户设备、便携式终端、膝上型终端、台式终端等其他名称。

通常，终端设备 600 包括有：处理器 601 和存储器 602。

处理器 601 可以包括一个或多个处理核心，比如 4 核心处理器、8 核心处理器等。处理器 601 可以采用 DSP（Digital Signal Processing，数字信号处理）、FPGA（Field-Programmable Gate Array，现场可编程门阵列）、PLA（Programmable Logic Array，可编程逻辑阵列）中的至少一种硬件形式来实现。处理器 601 也可以包括主处理器和协处理器，主处理器是用于对在唤醒状态下的数据进行处理的处理器，也称 CPU（Central Processing Unit，中央处理器）；协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中，处理器 601 可以集成有 GPU（Graphics Processing Unit，图像处理器），GPU 用于负责显示屏所需要显示的内容的渲染和绘制。一些实施例中，处理器 601 还可以包括 AI（Artificial Intelligence，人工智能）处理器，该 AI 处理器用于处理有关机器学习的计算操作。

存储器 602 可以包括一个或多个计算机可读存储介质，该计算机可读存储介质可以是非暂态的。存储器 602 还可包括高速随机存取存储器，以及非易失性存储器，比如一个或多个磁盘存储设备、闪存存储设备。在一些实施例中，存储器 602 中的非暂态的计算机可读存储介质用于存储至少一个指令，该至少一个指令用于被处理器 601 所执行以实现本申请实施例提供的音频处理方法。

在一些实施例中，终端设备 600 还可选包括有：外围设备接口 603 和至少一个外围设备。处理器 601、存储器 602 和外围设备接口 603 之间可以通过总线或信号线相连。各个外围设备可以通过总线、信号线或电路板与外围设备接口 603 相连。具体地，外围设备包括：射频电路 604、显示屏 605、摄像头组件 606、音频电路 607 和电源 608 中的至少一种。

外围设备接口 603 可被用于将 I/O（Input/Output，输入/输出）相关的至少一个外围设备连接到处理器 601 和存储器 602。在一些实施例中，处理器 601、存储器 602 和外围设备接口 603 被集成在同一芯片或电路板上；在一些其他实施例中，处理器 601、存储器 602 和外围设备接口 603 中的任意一个或两个可以在单独的芯片或电路板上实现，本实施例对此不加以限

定。

射频电路 604 用于接收和发射 RF（Radio Frequency，射频）信号，也称电磁信号。射频电路 604 通过电磁信号与通信网络以及其他通信设备进行通信。射频电路 604 将电信号转换为电磁信号进行发送，或者，将接收到的电磁信号转换为电信号。可选地，射频电路 604 包括：天线系统、RF 收发器、一个或多个放大器、调谐器、振荡器、数字信号处理器、编解码芯片组、用户身份模块卡等等。射频电路 604 可以通过至少一种无线通信协议来与其它终端设备进行通信。该无线通信协议包括但不限于：万维网、城域网、内联网、各代移动通信网络（2G、3G、4G 及 5G）、无线局域网和/或 WiFi（Wireless Fidelity，无线保真）网络。在一些实施例中，射频电路 604 还可以包括 NFC（Near Field Communication，近距离无线通信）有关的电路，本申请对此不加以限定。

显示屏 605 用于显示 UI（User Interface，用户界面）。该 UI 可以包括图形、文本、图标、视频及其它们的任意组合。当显示屏 605 是触摸显示屏时，显示屏 605 还具有采集在显示屏 605 的表面或表面上方的触摸信号的能力。该触摸信号可以作为控制信号输入至处理器 601 进行处理。此时，显示屏 605 还可以用于提供虚拟按钮和/或虚拟键盘，也称软按钮和/或软键盘。在一些实施例中，显示屏 605 可以为一个，设置在终端设备 600 的前面板；在另一些实施例中，显示屏 605 可以为至少两个，分别设置在终端设备 600 的不同表面或呈折叠设计；在另一些实施例中，显示屏 605 可以是柔性显示屏，设置在终端设备 600 的弯曲表面上或折叠面上。甚至，显示屏 605 还可以设置成非矩形的不规则图形，也即异形屏。显示屏 605 可以采用 LCD（Liquid Crystal Display，液晶显示屏）、OLED（Organic Light-Emitting Diode，有机发光二极管）等材质制备。

摄像头组件 606 用于采集图像或视频。可选地，摄像头组件 606 包括前置摄像头和后置摄像头。通常，前置摄像头设置在终端设备 600 的前面板，后置摄像头设置在终端设备 600 的背面。在一些实施例中，后置摄像头为至少两个，分别为主摄像头、景深摄像头、广角摄像头、长焦摄像头中的任意一种，以实现主摄像头和景深摄像头融合实现背景虚化功能、主摄像头和广角摄像头融合实现全景拍摄以及 VR（Virtual Reality，虚拟现实）拍摄功能或者其它融合拍摄功能。在一些实施例中，摄像头组件 606 还可以包括闪光灯。闪光灯可以是单色温闪光灯，也可以是双色温闪光灯。双色温闪光灯是指暖光闪光灯和冷光闪光灯的组合，可以用于不同色温下的光线补偿。

音频电路 607 可以包括麦克风和扬声器。麦克风用于采集用户及环境的声波，并将声波转换为电信号输入至处理器 601 进行处理，或者输入至射频电路 604 以实现语音通信。出于立体声采集或降噪的目的，麦克风可以为多个，分别设置在终端设备 600 的不同部位。麦克风还可以是阵列麦克风或全向采集型麦克风。扬声器则用于将来自处理器 601 或射频电路 604 的电信号转换为声波。扬声器可以是传统的薄膜扬声器，也可以是压电陶瓷扬声器。当扬声器是压电陶瓷扬声器时，不仅可以将电信号转换为人类可听见的声波，也可以将电信号转换为人类听不见的声波以进行测距等用途。在一些实施例中，音频电路 607 还可以包括耳机插孔。

电源 608 用于为终端设备 600 中的各个组件进行供电。电源 608 可以是交流电、直流电、一次性电池或可充电电池。当电源 608 包括可充电电池时，该可充电电池可以是有线充电电池或无线充电电池。有线充电电池是通过有线线路充电的电池，无线充电电池是通过无线线圈充电的电池。该可充电电池还可以用于支持快充技术。

在一些实施例中，终端设备 600 还包括有一个或多个传感器 609。该一个或多个传感器 609 包括但不限于：加速度传感器 610、陀螺仪传感器 611、压力传感器 612、光学传感器 613 以及接近传感器 614。

加速度传感器 610 可以检测以终端设备 600 建立的坐标系的三个坐标轴上的加速度大小。比如，加速度传感器 610 可以用于检测重力加速度在三个坐标轴上的分量。处理器 601 可以根据加速度传感器 610 采集的重力加速度信号，控制显示屏 605 以横向视图或纵向视图进行

用户界面的显示。加速度传感器 610 还可以用于游戏或者用户的运动数据的采集。

陀螺仪传感器 611 可以检测终端设备 600 的机体方向及转动角度，陀螺仪传感器 611 可以与加速度传感器 610 协同采集用户对终端设备 600 的 3D 动作。处理器 601 根据陀螺仪传感器 611 采集的数据，可以实现如下功能：动作感应（比如根据用户的倾斜操作来改变 UI）、拍摄时的图像稳定、游戏控制以及惯性导航。

压力传感器 612 可以设置在终端设备 600 的侧边框和/或显示屏 605 的下层。当压力传感器 612 设置在终端设备 600 的侧边框时，可以检测用户对终端设备 600 的握持信号，由处理器 601 根据压力传感器 612 采集的握持信号进行左右手识别或快捷操作。当压力传感器 612 设置在显示屏 605 的下层时，由处理器 601 根据用户对显示屏 605 的压力操作，实现对 UI 界面上的可操作性控件进行控制。可操作性控件包括按钮控件、滚动条控件、图标控件、菜单控件中的至少一种。

光学传感器 613 用于采集环境光强度。在一个实施例中，处理器 601 可以根据光学传感器 613 采集的环境光强度，控制显示屏 605 的显示亮度。具体地，当环境光强度较高时，调高显示屏 605 的显示亮度；当环境光强度较低时，调低显示屏 605 的显示亮度。在另一个实施例中，处理器 601 还可以根据光学传感器 613 采集的环境光强度，动态调整摄像头组件 606 的拍摄参数。

接近传感器 614，也称距离传感器，通常设置在终端设备 600 的前面板。接近传感器 614 用于采集用户与终端设备 600 的正面之间的距离。在一个实施例中，当接近传感器 614 检测到用户与终端设备 600 的正面之间的距离逐渐变小时，由处理器 601 控制显示屏 605 从亮屏状态切换为息屏状态；当接近传感器 614 检测到用户与终端设备 600 的正面之间的距离逐渐变大时，由处理器 601 控制显示屏 605 从息屏状态切换为亮屏状态。

本领域技术人员可以理解，图 6 中示出的结构并不构成对终端设备 600 的限定，可以包括比图示更多或更少的组件，或者组合某些组件，或者采用不同的组件布置。

图 7 为本申请实施例提供的服务器的结构示意图，该服务器 700 可因配置或性能不同而产生比较大的差异，可以包括一个或多个中央处理器（Central Processing Units，CPU）701 和一个或多个的存储器 702，其中，该一个或多个存储器 702 中存储有至少一条程序代码，该至少一条程序代码由该一个或多个处理器 701 加载并执行以实现上述各个方法实施例提供的音频处理方法。当然，该服务器 700 还可以具有有线或无线网络接口、键盘以及输入输出接口等部件，以便进行输入输出，该服务器 700 还可以包括其他用于实现设备功能的部件，在此不做赘述。

在示例性实施例中，还提供了一种计算机可读存储介质，该存储介质中存储有至少一条程序代码，该至少一条程序代码由处理器加载并执行，以使计算机实现上述任一种音频处理方法。

可选地，上述计算机可读存储介质可以是只读存储器（Read-Only Memory，ROM）、随机存取存储器（Random Access Memory，RAM）、只读光盘（Compact Disc Read-Only Memory，CD-ROM）、磁带、软盘和光数据存储设备等。

在示例性实施例中，还提供了一种计算机程序或计算机程序产品，该计算机程序或计算机程序产品中存储有至少一条计算机指令，该至少一条计算机指令由处理器加载并执行，以使计算机实现上述任一种音频处理方法。

需要说明的是，本申请所涉及的信息（包括但不限于用户设备信息、用户个人信息等）、数据（包括但不限于用于分析的数据、存储的数据、展示的数据等）以及信号，均为经用户授权或者经过各方充分授权的，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。例如，本申请中涉及到的音频片段都是在充分授权的情况下获取的。

权 利 要 求 书

1. 一种音频处理方法，由计算机设备执行，所述方法包括：

确定多个音频片段分别对应的声纹向量，所述声纹向量用于表示所述音频片段对应的声纹特征；

根据各个音频片段对应的声纹向量，确定初始相似度矩阵，所述初始相似度矩阵中包括任意两个音频片段对应的声纹向量之间的相似度；

根据所述初始相似度矩阵中各行对应的动态阈值，对所述初始相似度矩阵进行调整，得到参考相似度矩阵，所述动态阈值用于对不同相似度之间的相似度差值进行调节；

根据所述参考相似度矩阵确定发声得到所述多个音频片段的音频对象的数目；

根据所述音频对象的数目，对所述多个音频片段进行聚类，得到各个音频对象发声得到的音频片段。

2. 根据权利要求 1 所述的方法，其中，所述根据所述初始相似度矩阵中各行对应的动态阈值，对所述初始相似度矩阵进行调整，得到参考相似度矩阵之前，所述方法还包括：

对于所述初始相似度矩阵中的任一行，按照第一顺序对所述任一行中位于预设相似度范围内的相似度进行排序，得到第一排序结果；

确定所述第一排序结果中相邻的两个相似度之间的相似度差值，得到多个相似度差值；

在所述多个相似度差值中确定满足第一要求的相似度差值；

根据所述满足第一要求的相似度差值，确定所述任一行对应的动态阈值。

3. 根据权利要求 1 或 2 所述的方法，其中，所述根据所述初始相似度矩阵中各行对应的动态阈值，对所述初始相似度矩阵进行调整，得到参考相似度矩阵，包括：

将所述初始相似度矩阵第 k 行包括的相似度中，小于第 k 行对应的动态阈值的相似度调整为第一数值，并基于各行的调整结果得到所述参考相似度矩阵，k 为正整数；

或者，将所述初始相似度矩阵第 k 行包括的相似度中，小于所述第 k 行对应的动态阈值的相似度与第二数值相乘，并基于各行的调整结果得到所述参考相似度矩阵。

4. 根据权利要求 1 至 3 任一所述的方法，其中，所述根据所述参考相似度矩阵确定发声得到所述多个音频片段的音频对象的数目，包括：

根据多个参考参数，对所述参考相似度矩阵进行处理，得到各个参考参数对应的相似度矩阵；

根据所述多个参考参数和所述各个参考参数对应的相似度矩阵，确定所述多个音频片段中存在的音频对象的数目。

5. 根据权利要求 4 所述的方法，其中，所述根据多个参考参数，对所述参考相似度矩阵进行处理，得到各个参考参数对应的相似度矩阵，包括：

对于所述多个参考参数中的任一参考参数，根据所述任一参考参数，对所述参考相似度矩阵进行数值调整，得到第一相似度矩阵，所述数值调整用于简化所述参考相似度矩阵；

对所述第一相似度矩阵进行对称化处理，得到第二相似度矩阵，所述第二相似度矩阵中位于第 i 行第 j 列的相似度与位于第 j 行第 i 列的相似度相同，所述 i 和所述 j 为不大于所述多个音频片段的个数的正整数；

对所述第二相似度矩阵进行行列扩散，得到第三相似度矩阵，所述第三相似度矩阵用于生成多个音频对象之间的边界；

对所述第三相似度矩阵进行比例调整，得到第四相似度矩阵，所述比例调整用于将所述

第三相似度矩阵中各行包括的相似度调整在同一个范围内；

对所述第四相似度矩阵进行对称化处理，得到所述任一参考参数对应的相似度矩阵。

6. 根据权利要求 5 所述的方法，其中，所述根据所述任一参考参数，对所述参考相似度矩阵进行数值调整，得到第一相似度矩阵，包括：

对于所述参考相似度矩阵各行包括的多个相似度，将任一参考参数个满足第三要求的相似度之外的相似度调整为第三数值，得到所述第一相似度矩阵；

或者，将所述参考相似度矩阵包括的多个相似度中，除任一参考参数个满足第三要求的相似度之外的相似度与第四数值相乘，得到所述第一相似度矩阵。

7. 根据权利要求 5 或 6 所述的方法，其中，所述对所述第一相似度矩阵进行对称化处理，得到第二相似度矩阵，包括：

确定所述第一相似度矩阵对应的转置矩阵；

将所述第一相似度矩阵和所述第一相似度矩阵对应的转置矩阵中位于相同位置的相似度相加，得到待调整相似度矩阵；

对所述待调整相似度矩阵包括的多个相似度进行取半操作，得到所述第二相似度矩阵。

8. 根据权利要求 5 或 6 所述的方法，其中，所述对所述第一相似度矩阵进行对称化处理，得到第二相似度矩阵，包括：

确定所述第一相似度矩阵中位于所述第 i 行第 j 列的相似度，与所述第一相似度矩阵中位于所述第 j 行第 i 列的相似度中最大的相似度，将所述最大的相似度作为所述第二相似度矩阵中位于所述第 i 行第 j 列和所述第 j 行第 i 列的相似度，得到所述第二相似度矩阵。

9. 根据权利要求 5 至 8 任一所述的方法，其中，所述对所述第二相似度矩阵进行行列扩散，得到第三相似度矩阵，包括：

确定所述第二相似度矩阵对应的转置矩阵；

根据所述第二相似度矩阵和所述第二相似度矩阵对应的转置矩阵，确定所述第三相似度矩阵，所述第三相似度矩阵中位于第 m 行第 n 列的相似度基于所述第二相似度矩阵中位于所述第 m 行的相似度和所述第二相似度矩阵对应的转置矩阵中位于所述第 n 列的相似度确定，所述 m、所述 n 为不大于所述多个音频片段的个数的正整数。

10. 根据权利要求 5 至 9 任一所述的方法，其中，所述对所述第三相似度矩阵进行比例调整，得到第四相似度矩阵，包括：

根据所述第三相似度矩阵中各行包括的多个相似度，确定各行对应的最大相似度；

将所述第三相似度矩阵中各行包括的多个相似度分别与所述各行对应的最大相似度相除，得到所述第四相似度矩阵。

11. 根据权利要求 4 至 10 任一所述的方法，其中，所述根据所述多个参考参数和所述各个参考参数对应的相似度矩阵，确定所述多个音频片段中存在的音频对象的数目，包括：

根据所述多个参考参数和所述各个参考参数对应的相似度矩阵，确定所述各个参考参数对应的比例值，所述比例值用于指示所述参考参数对应的相似度矩阵中保留的相似度的数量；

根据所述各个参考参数对应的比例值，确定所述多个音频片段中存在的音频对象的数目。

12. 根据权利要求 11 所述的方法，其中，所述根据所述多个参考参数和所述各个参考参数对应的相似度矩阵，确定所述各个参考参数对应的比例值，包括：

对于所述多个参考参数中的任一参考参数，对所述任一参考参数对应的相似度矩阵进行

拉普拉斯变换，得到所述任一参考参数对应的拉普拉斯矩阵；

对所述拉普拉斯矩阵进行奇异值分解，得到多个参考特征值；

在所述多个参考特征值中确定第二特征值和第一数量个第一特征值，所述第二特征值为所述多个参考特征值中的最大值，所述第一特征值为按照第二顺序对所述多个参考特征值进行排序后满足第二要求的参考特征值；

确定所述第一数量个第一特征值中相邻的两个第一特征值之间的差值，得到多个特征值差值；

根据所述第二特征值，对第一特征值差值进行归一化处理，得到归一化之后的特征值差值，所述第一特征值差值为所述多个特征值差值中最大的特征值差值；

根据所述归一化之后的特征值差值和所述任一参考参数，确定所述任一参考参数对应的比例值。

13. 根据权利要求 11 或 12 所述的方法，其中，所述根据所述各个参考参数对应的比例值，确定所述多个音频片段中存在的音频对象的数目，包括：

根据所述各个参考参数对应的比例值，在所述多个参考参数中确定第一参数，所述第一参数为所述多个参考参数中对应的比例值最小的参考参数；

确定所述第一参数对应的多个特征值差值；

调用第一函数对所述第一参数对应的多个特征值差值进行处理，得到所述多个音频片段中存在的音频对象的数目。

14. 根据权利要求 13 所述的方法，其中，所述根据所述音频对象的数目，对所述多个音频片段进行聚类，得到各个音频对象发声得到的音频片段，包括：

对所述第一参数对应的相似度矩阵进行奇异值分解，得到多个分解特征值；

在所述多个分解特征值中确定与所述音频对象的数目对应分解特征值；

确定所述音频对象的数目个分解特征值分别对应的特征向量，并生成分解矩阵，所述分解矩阵的行数为所述音频对象的数目，列数为所述音频片段的数目；

根据所述分解矩阵，确定所述多个音频片段分别对应的特征向量，所述特征向量用于指示对应的音频片段；

根据所述音频对象的数目和所述多个音频片段分别对应的特征向量，对所述多个音频片段进行聚类，各个音频对象发声得到的音频片段。

15. 一种音频处理装置，所述装置包括：

确定模块，用于确定多个音频片段分别对应的声纹向量，所述声纹向量用于表示所述音频片段对应的声纹特征；

所述确定模块，还用于根据各个音频片段对应的声纹向量，确定初始相似度矩阵，所述初始相似度矩阵中包括任意两个音频片段对应的声纹向量之间的相似度；

调整模块，用于根据所述初始相似度矩阵中各行对应的动态阈值，对所述初始相似度矩阵进行调整，得到参考相似度矩阵，所述动态阈值用于对不同相似度之间的相似度差值进行调节；

所述确定模块，还用于根据所述参考相似度矩阵确定发声得到所述多个音频片段的音频对象的数目；

聚类模块，用于根据所述音频对象的数目，对所述多个音频片段进行聚类，得到各个音频对象发声得到的音频片段。

16. 根据权利要求 15 所述的装置，其中，所述确定模块，还用于对于所述初始相似度矩阵中的任一行，按照第一顺序对所述任一行中位于预设相似度范围内的相似度进行排序，得

到第一排序结果；确定所述第一排序结果中相邻的两个相似度之间的相似度差值，得到多个相似度差值；在所述多个相似度差值中确定满足第一要求的相似度差值；根据所述满足第一要求的相似度差值，确定所述任一行对应的动态阈值。

17. 根据权利要求 15 或 16 所述的装置，所述调整模块，用于将所述初始相似度矩阵第 k 行包括的相似度中，小于第 k 行对应的动态阈值的相似度调整为第一数值，并基于各行的调整结果得到所述参考相似度矩阵，k 为正整数；或者，将所述初始相似度矩阵第 k 行包括的相似度中，小于所述第 k 行对应的动态阈值的相似度与第二数值相乘，并基于各行的调整结果得到所述参考相似度矩阵。

18. 一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条程序代码，所述至少一条程序代码由所述处理器加载并执行，以使所述计算机设备实现如权利要求 1 至 14 任一所述的音频处理方法。

19. 一种计算机可读存储介质，所述计算机可读存储介质中存储有至少一条程序代码，所述至少一条程序代码由处理器加载并执行，以使计算机实现如权利要求 1 至 14 任一所述的音频处理方法。

20. 一种计算机程序产品，包括计算机程序或指令，所述计算机程序或指令被处理器执行时实现如权利要求 1 至 14 任一所述的音频处理方法。

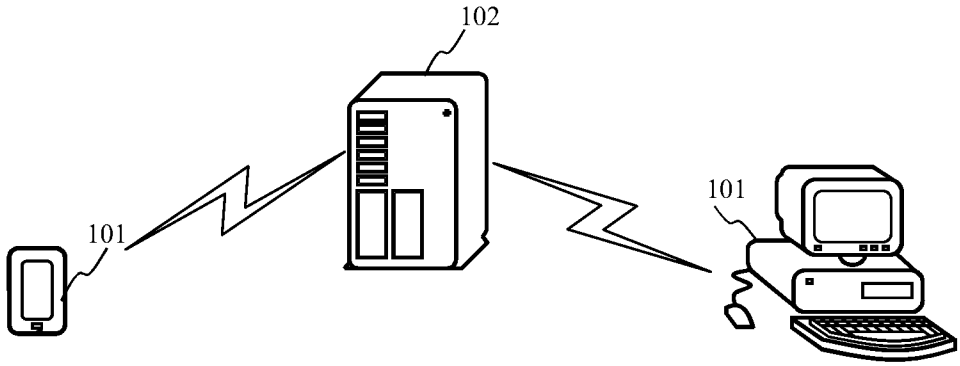
102
101
101

图 1

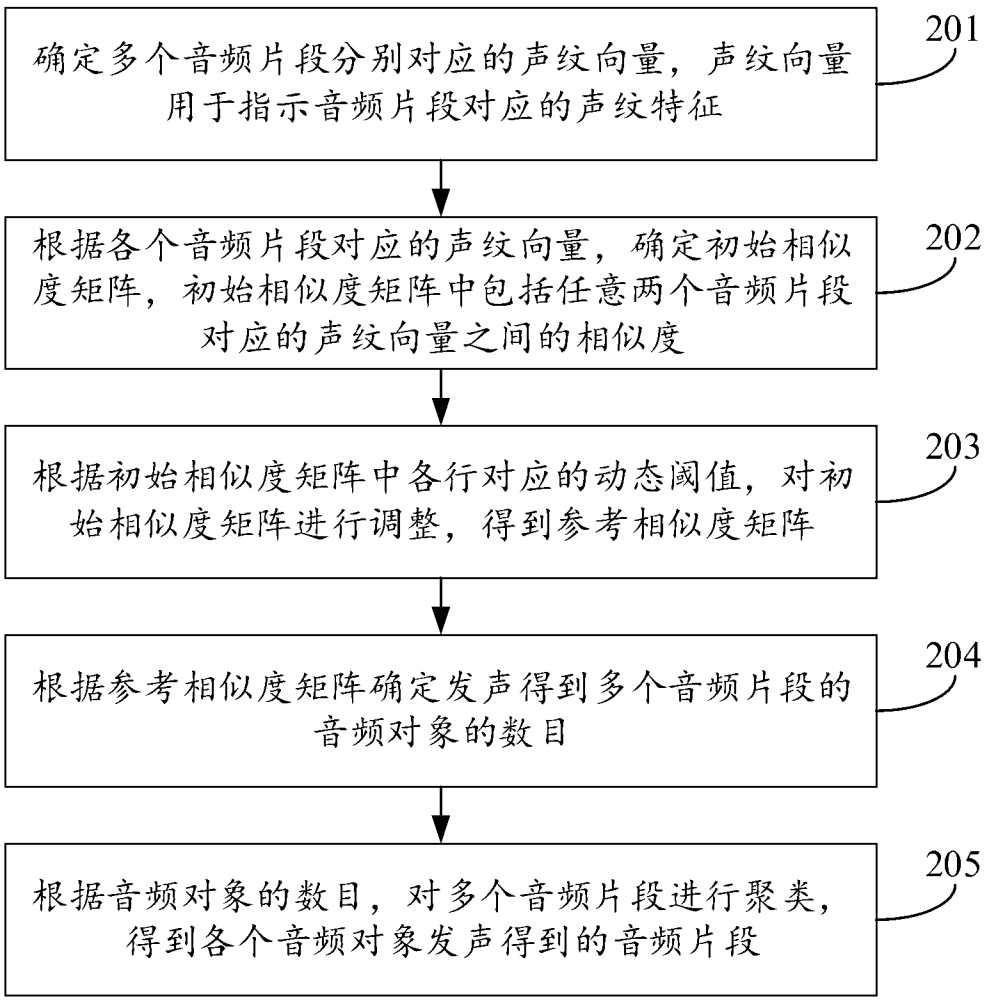
确定多个音频片段分别对应的声纹向量，声纹向量用于指示音频片段对应的声纹特征
201
根据各个音频片段对应的声纹向量，确定初始相似度矩阵，初始相似度矩阵中包括任意两个音频片段对应的声纹向量之间的相似度
202
根据初始相似度矩阵中各行对应的动态阈值，对初始相似度矩阵进行调整，得到参考相似度矩阵
203
根据参考相似度矩阵确定发声得到多个音频片段的音频对象的数目
204
根据音频对象的数目，对多个音频片段进行聚类，得到各个音频对象发声得到的音频片段
205

图 2

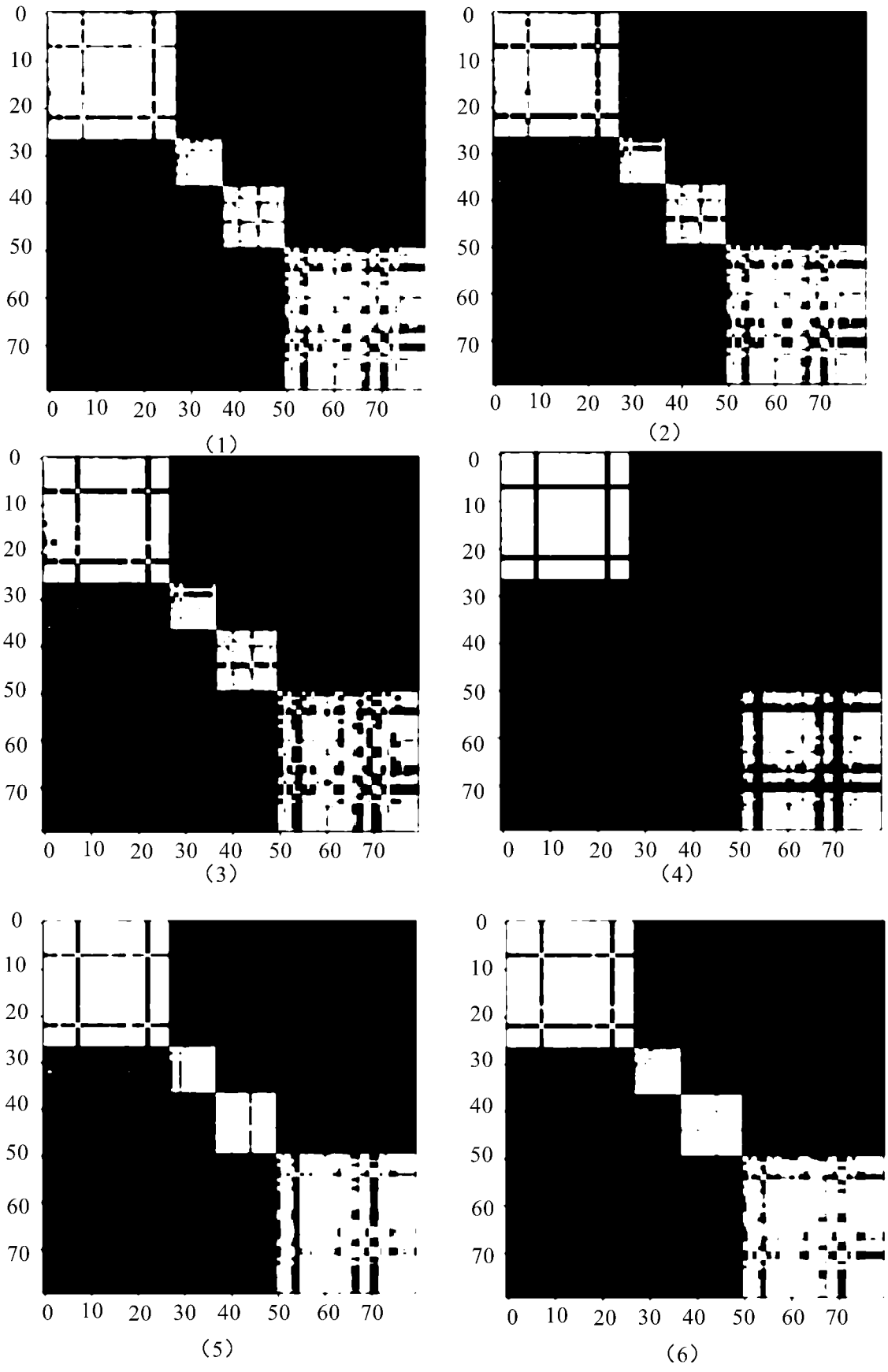

图 3

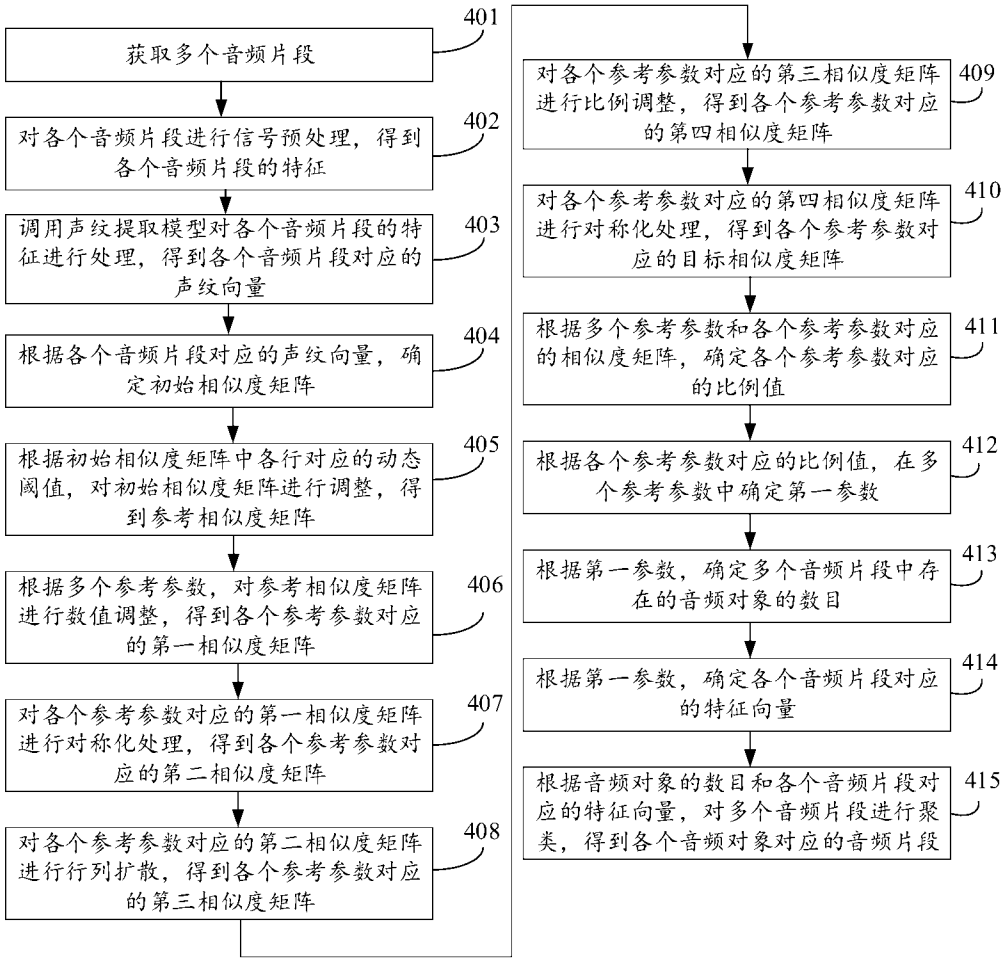
401
获取多个音频片段
402
对各个音频片段进行信号预处理，得到各个音频片段的特征
403
调用声纹提取模型对各个音频片段的特征进行处理，得到各个音频片段对应的声纹向量
404
根据各个音频片段对应的声纹向量，确定初始相似度矩阵
405
根据初始相似度矩阵中各行对应的动态阈值，对初始相似度矩阵进行调整，得到参考相似度矩阵
406
根据多个参考参数，对参考相似度矩阵进行数值调整，得到各个参考参数对应的第一相似度矩阵
407
对各个参考参数对应的第一相似度矩阵进行对称化处理，得到各个参考参数对应的第二相似度矩阵
408
对各个参考参数对应的第二相似度矩阵进行行列扩散，得到各个参考参数对应的第三相似度矩阵
409
对各个参考参数对应的第三相似度矩阵进行比例调整，得到各个参考参数对应的第四相似度矩阵
410
对各个参考参数对应的第四相似度矩阵进行对称化处理，得到各个参考参数对应的目标相似度矩阵
411
根据多个参考参数和各个参考参数对应的相似度矩阵，确定各个参考参数对应的比例值
412
根据各个参考参数对应的比例值，在多个参考参数中确定第一参数
413
根据第一参数，确定多个音频片段中存在的音频对象的数目
414
根据第一参数，确定各个音频片段对应的特征向量
415
根据音频对象的数目和各个音频片段对应的特征向量，对多个音频片段进行聚类，得到各个音频对象对应的音频片段

图 4

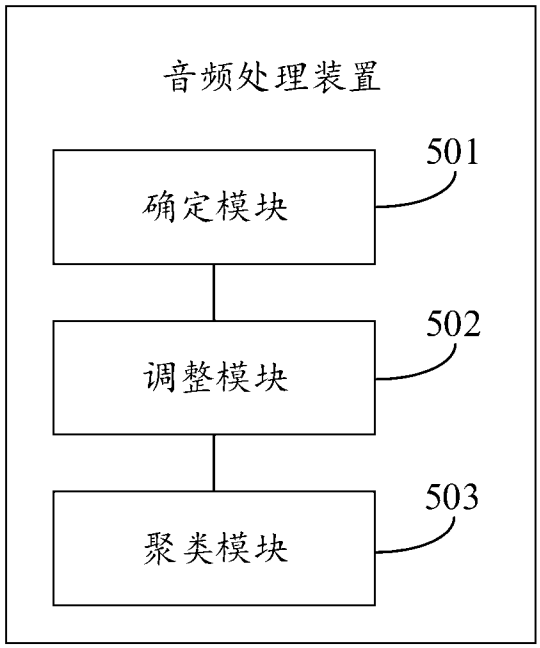
音频处理装置
501
确定模块
502
调整模块
503
聚类模块

图 5

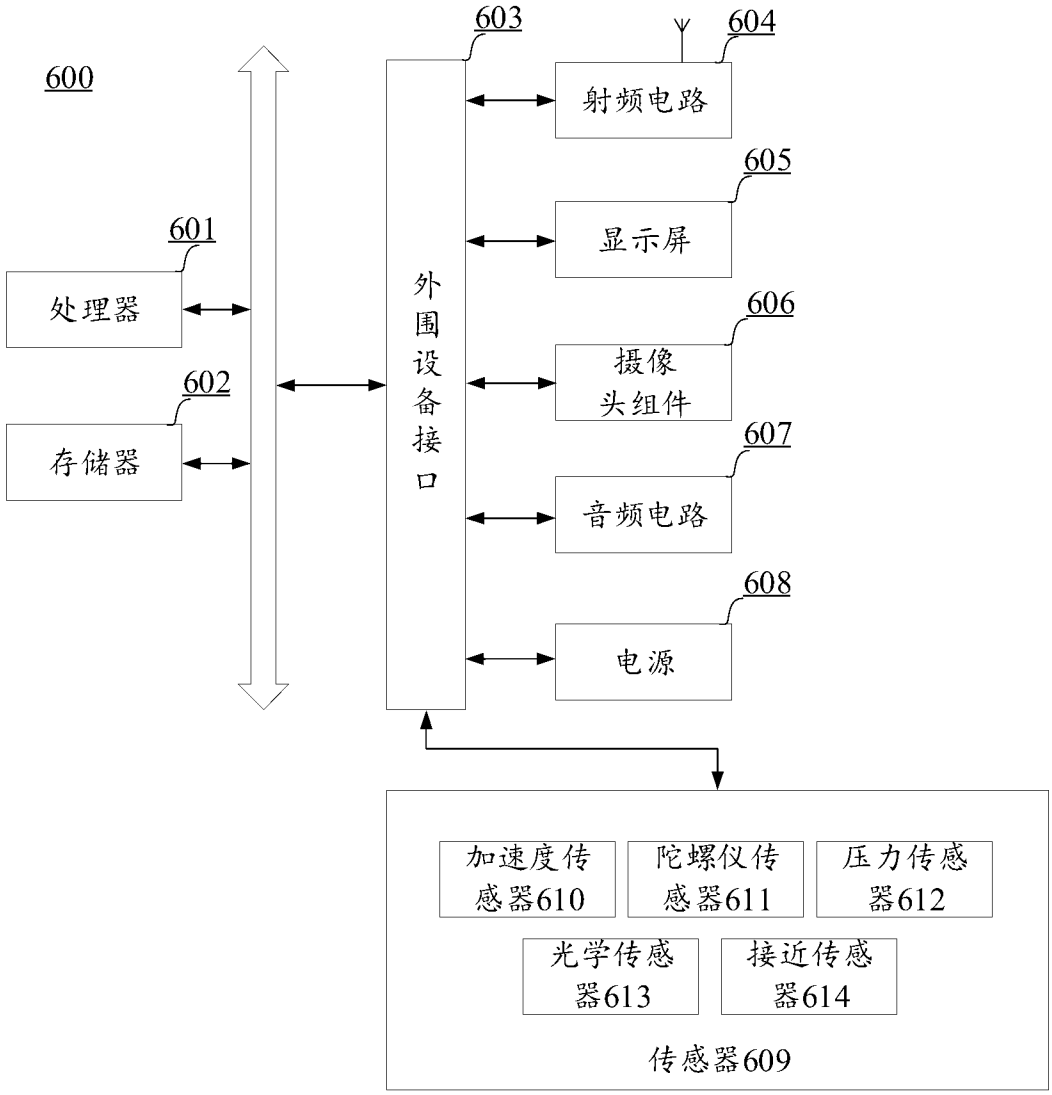
600
601
处理器
602
存储器
外围设备接口
603
604
射频电路
605
显示屏
606
摄像头组件
607
音频电路
608
电源
加速度传感器610
陀螺仪传感器611
压力传感器612
光学传感器613
接近传感器614
传感器609

图 6

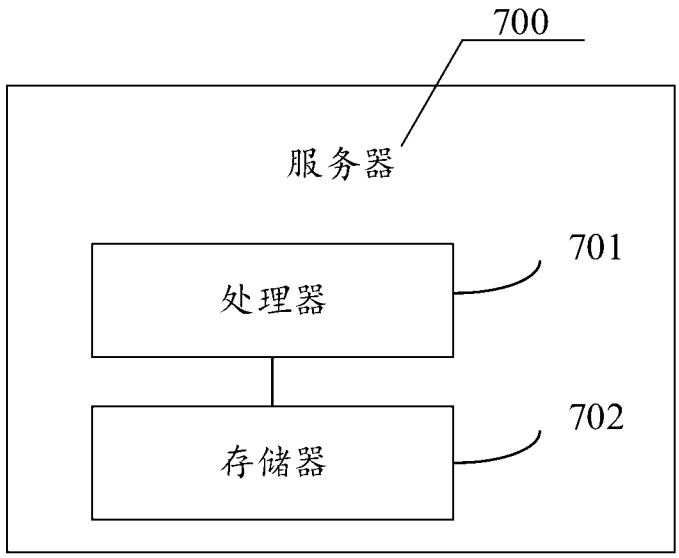
700
服务器
701
处理器
702
存储器

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/114040

A. CLASSIFICATION OF SUBJECT MATTER

G06F16/65(2019.01)i; G10L25/27(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06F G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 音频, 声纹, 向量, 相似度, 矩阵, 初始, 调整, 动态阈值, 数目, 聚类, audio, voiceprint, vector, similarity, matrix, initial, adjust, dynamic threshold, number, cluster

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 115168643 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 11 October 2022 (2022-10-11) claims 1-17, and description, paragraphs [0007]-[0023]	1-20
X	CN 114446284 A (SHANGHAI XIMALAYA TECHNOLOGY CO., LTD.) 06 May 2022 (2022-05-06) description, paragraphs [0004]-[0036]	1-4, 15-20
A	CN 113327628 A (BEIJING BYTEDANCE NETWORK TECHNOLOGY CO., LTD.) 31 August 2021 (2021-08-31) entire document	1-20
A	CN 114822558 A (MASHANG CONSUMER FINANCE CO., LTD.) 29 July 2022 (2022-07-29) entire document	1-20
A	WO 2021072893 A1 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 22 April 2021 (2021-04-22) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
03 November 2023	**10 November 2023**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/114040

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 115168643 A	11 October 2022	None	
CN 114446284 A	06 May 2022	None	
CN 113327628 A	31 August 2021	None	
CN 114822558 A	29 July 2022	None	
WO 2021072893 A1	22 April 2021	CN 110889009 A	07 March 2020

国际检索报告

国际申请号
PCT/CN2023/114040

A. 主题的分类

G06F16/65(2019.01)i；G10L25/27(2013.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC：G06F G10L

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

VEN, CNABS, CNTXT, WOTXT, EPTXT, USTXT, CNKI, IEEE: 音频，声纹，向量，相似度，矩阵，初始，调整，动态阈值，数目，聚类，audio, voiceprint, vector, similarity, matrix, initial, adjust, dynamic threshold, number, cluster

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 115168643 A (腾讯科技（深圳）有限公司) 2022年10月11日 (2022 - 10 - 11) 权利要求1−17, 说明书第[0007]−[0023]段	1−20
X	CN 114446284 A (上海喜马拉雅科技有限公司) 2022年5月6日 (2022 - 05 - 06) 说明书第[0004]−[0036]段	1−4,15−20
A	CN 113327628 A (北京字节跳动网络技术有限公司) 2021年8月31日 (2021 - 08 - 31) 全文	1−20
A	CN 114822558 A (马上消费金融股份有限公司) 2022年7月29日 (2022 - 07 - 29) 全文	1−20
A	WO 2021072893 A1 (PING AN TECHNOLOGY(SHENZHEN) CO., LTD.) 2021年4月22日 (2021 - 04 - 22) 全文	1−20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“D” 申请人在国际申请中引证的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2023年11月3日	2023年11月10日
ISA/CN的名称和邮寄地址 中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	受权官员 郭婉莹 电话号码 (+86) 010−53961361

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/114040

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 115168643 A	2022年10月11日	无	
CN 114446284 A	2022年5月6日	无	
CN 113327628 A	2021年8月31日	无	
CN 114822558 A	2022年7月29日	无	
WO 2021072893 A1	2021年4月22日	CN 110889009 A	2020年3月7日