

(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 1 月 4 日 (04.01.2024)



(10) 国际公布号
WO 2024/001646 A1

(51) 国际专利分类号:
G10L 15/02 (2006.01) **G06N 3/04** (2023.01)
G10L 15/16 (2006.01)

(21) 国际申请号: PCT/CN2023/097205

(22) 国际申请日: 2023 年 5 月 30 日 (30.05.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202210747175.3 2022年6月29日 (29.06.2022) 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 518000 (CN)。

(72) 发明人: 冯鑫 (FENG, Xin); 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 518000 (CN)。

(74) 代理人: 北京派特恩知识产权代理有限公司 (CHINA PAT INTELLECTUAL PROPERTY OFFICE); 中国北京市海淀区苏州街 3 号大恒科技大厦南座五层 503, Beijing 100080 (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, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD,

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

(54) 发明名称: 音频数据的处理方法、装置、电子设备、程序产品及存储介质

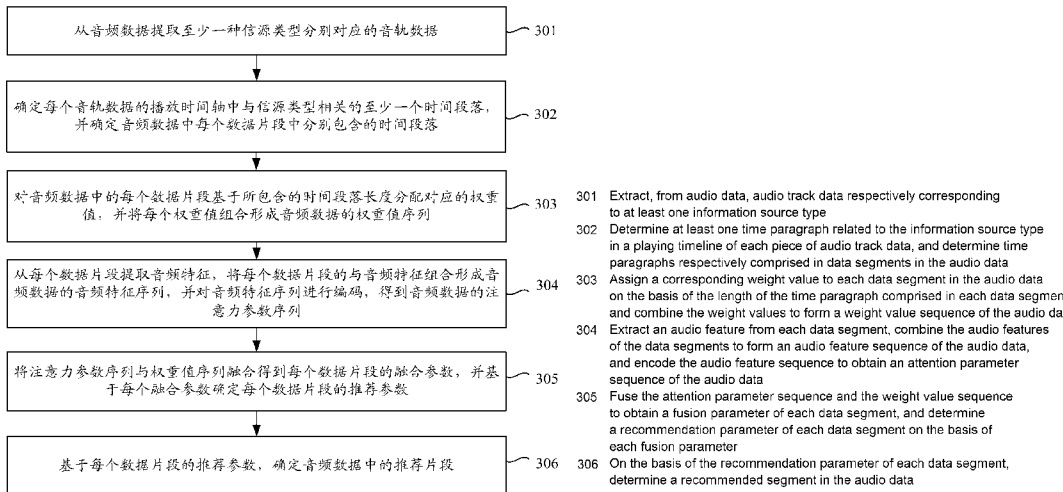


图 3A

(57) Abstract: An audio data processing method and apparatus, an electronic device, a program product, and a storage medium. The method comprises: extracting, from audio data, audio track data respectively corresponding to at least one information source type (301); determining at least one time paragraph related to the information source type in a playing timeline of each piece of audio track data, and determining time paragraphs respectively comprised in data segments in the audio data (302); assigning a corresponding weight value to each data segment in the audio data, and combining the weight values to form a weight value sequence of the audio data (303); extracting an audio feature from each data segment, combining the audio features of the data segments to form an audio

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))。

feature sequence of the audio data, and encoding the audio feature sequence to obtain an attention parameter sequence of the audio data (304); fusing the attention parameter sequence and the weight value sequence to obtain a fusion parameter of each data segment, and determining a recommendation parameter of each data segment on the basis of each fusion parameter (305); and on the basis of each recommendation parameter, determining a recommended data segment in the audio data (306).

(57) 摘要: 一种音频数据的处理方法、装置、电子设备、程序产品及存储介质。该方法包括: 从音频数据提取得到至少一种信源类型分别对应的音轨数据 (301); 确定每个音轨数据的播放时间轴中与信源类型相关的至少一个时间段落, 并确定音频数据中每个数据片段中分别包含的时间段落 (302); 对音频数据中的每个数据片段分配对应的权重值, 将每个权重值组合形成音频数据的权重值序列 (303); 从每个数据片段提取音频特征, 将每个数据片段的与音频特征组合形成音频数据的音频特征序列, 对音频特征序列进行编码, 得到音频数据的注意力参数序列 (304); 将注意力参数序列与权重值序列融合得到每个数据片段的融合参数, 并基于每个融合参数确定每个数据片段的推荐参数 (305); 基于每个推荐参数, 确定音频数据中的推荐数据片段 (306)。

音频数据的处理方法、装置、电子设备、程序产品及存储介质

相关申请的交叉引用

本申请基于申请号为 202210747175.3、申请日为 2022 年 6 月 29 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

本申请涉及计算机技术，尤其涉及一种音频数据的处理方法、装置、电子设备、程序产品及存储介质。

背景技术

在线多媒体（例如视频或音频）播放平台需要在多媒体数据中标记出一些特殊的数据片段，称为推荐片段，例如精彩数据片段、热门数据片段等，以方便用户观看。

相关技术通过视频/音频的播放记录数据确定视频/音频的推荐数据片段，但是针对新发布的视频/音频，没有播放记录数据，只能通过人工标注推荐片段的方式。例如：通过人工标注的方式来定位整个剧集视频的精彩数据片段。但是人工标注严重依赖于人工标注的主观感受，标注出的推荐片段根据标注人的差异而存在差异，并且人工标注的耗时过长，标注效率低，无法进行快速的批量化生产。

综上所述，针对没有大量播放记录数据的多媒体数据，暂无较好的方式识别推荐数据片段。

发明内容

本申请实施例提供一种音频数据的处理方法、装置、电子设备、计算机程序产品及计算机可读存储介质、计算机程序产品，能够从音频数据中准确识别出推荐片段。

本申请实施例的技术方案是这样实现的：

本申请实施例提供一种音频数据的处理方法，由电子设备执行，包括：

从音频数据提取得到至少一种信源类型分别对应的音轨数据，其中，所述音频数据包含多个数据片段；

确定每个所述音轨数据的播放时间轴中与所述信源类型相关的至少一个时间段落，并确定所述音频数据中每个所述数据片段中分别包含的时间段落；

对所述音频数据中的每个数据片段基于所包含的所述时间段落长度分配对应的权重值，并将每个所述权重值组合形成所述音频数据的权重值序列；

从所述每个数据片段提取音频特征，将所述每个数据片段的与音频特征组合形成所述音频数据的音频特征序列，并对所述音频特征序列进行编码，得到所述音频数据的注意力参数序列；

将所述注意力参数序列与所述权重值序列融合得到每个所述数据片段的融合参数，并基于每个所述融合参数确定每个所述数据片段的推荐参数；

基于每个所述数据片段的推荐参数，确定所述音频数据中的推荐片段。

本申请实施例提供一种音频数据的处理装置，包括：

信源分离模块，配置为从音频数据提取得到至少一种信源类型分别对应的音轨数据，

其中，所述音频数据包含多个数据片段；

权重配置模块，配置为确定每个所述音轨数据的播放时间轴中与所述信源类型相关的至少一个时间段落，并确定所述音频数据中每个所述数据片段中分别包含的时间段落；

对所述音频数据中的每个数据片段基于所包含的所述时间段落长度分配对应的权重值，并将每个所述权重值组合形成所述音频数据的权重值序列；

特征提取模块，配置为从所述每个数据片段提取音频特征，将所述每个数据片段的与音频特征组合形成所述音频数据的音频特征序列，并对所述音频特征序列进行编码，得到所述音频数据的注意力参数序列；

参数预测模块，配置为将所述注意力参数序列与所述权重值序列融合得到每个所述数据片段的融合参数，并基于每个所述融合参数确定每个所述数据片段的推荐参数；

所述参数预测模块，还配置为基于每个所述数据片段的推荐参数，确定所述音频数据中的推荐片段。

本申请实施例提供一种电子设备，包括：

存储器，用于存储可执行指令；

处理器，用于执行所述存储器中存储的可执行指令时，实现本申请实施例提供的音频数据的处理方法。

本申请实施例提供一种计算机可读存储介质，存储有可执行指令，用于引起处理器执行时，实现本申请实施例提供的音频数据的处理方法。

本申请实施例提供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本申请实施例提供的音频数据的处理方法。

本申请实施例具有以下有益效果：

一方面，通过对音频数据提取至少一种信源对应的音轨数据，基于信源相关的时间段落对数据片段分配对应的权重值，从而，在将各个数据片段的权重值组成的权重值序列与注意力参数序列融合时，能够从时域突出与信源对应的数据片段的重要程度；另一方面，通过注意力参数序列来从频域层面突出音频特征中与信源相关的数据片段的特征，这样，通过对音频数据的音轨的时域、频域两个方面的信息进行量化，预测音频数据中每个数据片段属于某一类型的数据片段的概率（推荐参数），相较于单纯从频域的层面来预测，识别更加全面，从而基于每个数据片段的推荐参数可以准确识别出有价值的推荐片段，为用户提供准确的参考信息。

附图说明

图1是本申请实施例提供的音频数据的处理方法的应用模式示意图；

图2是本申请实施例提供的电子设备的结构示意图；

图3A是本申请实施例提供的音频数据的处理方法的第一流程示意图；

图3B是本申请实施例提供的音频数据的处理方法的第二流程示意图；

图3C是本申请实施例提供的音频数据的处理方法的第三流程示意图；

图3D是本申请实施例提供的音频数据的处理方法的第四流程示意图；

图3E是本申请实施例提供的音频数据的处理方法的第五流程示意图；

图4A是本申请实施例提供的视频中提取的音频数据的示意图；

图4B是本申请实施例提供的音轨数据示意图；

图4C是本申请实施例提供的时间段落示意图；

图5是本申请实施例提供的音频数据的处理方法的一个可选的流程示意图；

图6A是本申请实施例提供的音频处理模型的第一示意图；

图6B是本申请实施例提供的音频处理模型的第二示意图；

图 7 是本申请实施例提供的金字塔场景解析网络的示意图；
图 8 是本申请实施例提供的音频语义信息提取模块的示意图；
图 9 是本申请实施例提供的注意力模块中编码的原理示意图；
图 10A 是本申请实施例提供的播放界面的第一示意图；
图 10B 是本申请实施例提供的播放界面的第二示意图；
图 11 是本申请实施例提供的音频数据的示意图。

具体实施方式

为了使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请作进一步地详细描述，所描述的实施例不应视为对本申请的限制，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例，都属于本申请保护的范围。

在以下的描述中，涉及到“一些实施例”，其描述了所有可能实施例的子集，但是可以理解，“一些实施例”可以是所有可能实施例的相同子集或不同子集，并且可以在不冲突的情况下相互结合。

在以下的描述中，所涉及的术语“第一\第二\第三”仅仅是区别类似的对象，不代表针对对象的特定排序，可以理解地，“第一\第二\第三”在允许的情况下可以互换特定的顺序或先后次序，以使这里描述的本申请实施例能够以除了在这里图示或描述的以外的顺序实施。

需要指出，在本申请实施例中，涉及到用户信息、用户反馈数据等相关的数据（例如：多媒体数据、语音、音轨数据等），当本申请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。

除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中所使用的术语只是为了描述本申请实施例的目的，不是旨在限制本申请。

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) 金字塔场景解析网络 (PSPN, Pyramid Scene Parsing Network)，金字塔场景解析网络的作用是预测所关注对象的标签 (label)、位置 (location) 和形状 (shape)。该网络中包括金字塔池化模块 (Pyramid Pooling Module)，金字塔池化模块可以将局部的上下文信息进行聚合，形成全局的上下文信息，更全面地实现定位、分类等处理。

2) 信源分离，在音频数据（例如从视频数据的音频轨道中提取的音频数据，或，从音频文件中提取的音频数据）中，可能会承载一种或多种音频信号（即数字音频信号的简称，数字音频信号是对模拟音频信号进行采样和编码得到的），信源是发出声音信号的来源，信源类型是发声来源的类型，每种音频信号对应一种信源类型（例如语音对应的信源类型为人类），而信源分离就是通过信号处理或者其他算法进行分离处理，提取出指定信源的音频信号的序列，最终生成由不同信源类型的音频信号的序列分别构成的音轨数据，例如：语音音轨数据，背景音轨数据。

3) 语音活动检测 (VAD, Voice Activity Detection) 算法，用于检测音频中语音/非语音（非语音/静音）的算法。广泛应用于语音编码、降噪和自动语音识别等场景 (ASR, Automatic Speech Recognition) 中。

4) 时域和频域，时域和频域是音频数据的基本性质，用来分析音频数据的不同角度称为域，是衡量音频特征的两个维度概念。时域维度下，将音频数据的采样点在时间上进行展示处理，信号与时间之间存在相应的关系。通过傅里叶变换可以把信号从时域

转换到频域。频域用于分析音频数据在各个频带上的能量分布，包含音频数据一定程度上的特征表现。

5 5) 梅尔 (Mel) 频率，一种基于人耳对等距的音高 (pitch) 变化的感官判断而定的非线性频率刻度，是在进行信号处理时，更能够迎合人耳的听觉感受阈值变化来人为设定的频率刻度，在音频处理领域，有很多基础音频特征是通过 mel 频率来进行计算的。

6) 卷积神经网络 (CNN, Convolutional Neural Networks)，是一类包含卷积计算且具有深度结构的前馈神经网络 (FNN, Feed forward Neural Networks)，是深度学习 (Deep Learning) 的代表算法之一。卷积神经网络具有表征学习 (Representation Learning) 能力，能够按其阶层结构对输入图像进行平移不变分类 (Shift-invariant Classification)。卷积神经网络的人工神经元可以响应一部分覆盖范围内的周围单元。卷积神经网络由一个或多个卷积层和顶端的全连接层 (对应经典的神经网络) 组成，同时也包括关联权重和池化层 (pooling layer)。

7) 注意力 (Attention) 机制，模仿人类注意力而提出的一种解决问题的办法，能够从大量信息中快速筛选出高价值信息。注意力机制用于主要用于解决长短期记忆网络 (LSTM, Long Short-Term Memory)、循环神经网络 (RNN, Recurrent Neural Network) 模型输入序列较长的时候很难获得最终合理的向量表示问题，做法是保留长短期记忆网络的中间结果，用新的模型对中间结果与输出结果之间的关联性进行学习，从而确定输出结果中精彩程度较高的信息，从而达到信息筛选的目的。

8) 时间段落，多媒体数据的播放时间轴上的一个区间，例如时长为 10 分钟的视频，在播放时间轴上从第 5 分钟至第 8 分钟的区间，可以称为 1 个时间段落。

9) 数据片段，多媒体数据中对应时间段落的数据。例如时长为 10 分钟的视频，在播放时间轴上从第 5 分钟至第 8 分钟的时间段落所对应的数据，可以称为 1 个数据片段，可以区分为音轨轨道的数据片段和视频轨道的数据片段。一个视频可以划分为多个时长相等的数据片段。

10) 推荐片段，多媒体数据中包括待表达的关键信息或极性情感 (例如悲伤、愉快) 的数据片段，在播放时间上对应播放时间轴上的一个时间数据片段，多媒体数据可以是视频、歌曲、有声小说和广播剧等，推荐片段可以是以下类型：电影中包括关键情节的精彩片段，歌曲中抒发悲伤情感的悲伤片段等。

11) 推荐参数，量化表征一个数据片段属于某个特定类型的推荐片段的概率，例如：推荐参数表征推荐片段是多媒体数据中的精彩片段的概率。

本申请实施例提供一种音频数据的处理方法、音频数据的处理装置、电子设备、计算机程序产品和计算机可读存储介质，能够准确获取音频数据中的推荐片段。

参考图 1，图 1 是本申请实施例提供的音频数据的处理方法的应用模式示意图；示例的，涉及的服务器包括：识别服务器 201 与媒体服务器 202，其中，媒体服务器 202 可以是视频平台的服务器、音乐平台的服务器、有声小说或者广播剧平台的服务器等。图 1 中还示出了网络 300 及终端设备 401。识别服务器 201 与媒体服务器 202 之间通过网络 300 进行通信，或者通过其他方式进行通信，终端设备 401 通过网络 300 连接媒体服务器 202，网络 300 可以是广域网或者局域网，又或者是二者的组合。

在一些实施例中，媒体服务器 202 将音频数据 (例如有声小说、在线音乐) 发送给识别服务器 201，识别服务器 201 确定音频数据中每个数据片段的推荐参数 (例如，数据片段属于精彩片段、悲伤片段、搞笑数据片段的概率，推荐参数与精彩程度、悲伤程度、搞笑程度等正相关)，并基于推荐参数生成推荐参数曲线、确定音频数据中的推荐片段。将推荐参数曲线与推荐片段发送至媒体服务器 202，媒体服务器 202 将推荐参数曲线、推荐片段位置标签发送给终端设备 401，终端设备 401 运行播放器 402，当播放

器 402 播放对应的音频数据时, 显示推荐参数曲线、推荐片段位置标签。用户基于推荐参数曲线能够方便地确定音频数据中每个数据片段的推荐参数, 以及基于推荐片段位置标签能够跳转到对应的位置进行播放, 便于定位推荐片段。

在一些实施例中, 从视频数据 (例如在线视频或本地视频) 的音频轨道中分割出音频数据, 得到音频数据。音频数据与视频画面的时间轴是对齐的, 音频数据的精彩数据片段与视频数据的精彩数据片段一一对应。推荐片段可以分别是精彩数据片段、悲伤数据片段、搞笑数据片段等。下面以推荐参数是数据片段属于精彩片段的概率, 推荐片段是精彩数据片段举例说明。用户可以是观看视频的观众, 或者将视频数据作为素材进行二次创作的用户。用户通过精彩数据片段的推荐参数曲线、位置标签可以快速确定视频中的精彩数据片段, 进而观看精彩数据片段, 或者将精彩数据片段从视频数据中剪切出来, 进行视频二次创作。

在一些实施例中, 识别服务器 201 与媒体服务器 202 可以集成在一起实施为一个统一的服务器, 也可以分开设置。

本申请实施例可以通过区块链技术实现, 可以将本申请实施例的音频数据的处理方法得到的推荐参数曲线为检测结果, 将检测结果上传到区块链中存储, 通过共识算法保证检测结果的可靠性。区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链 (Block chain), 本质上是一个去中心化的数据库, 是一串使用密码学方法相关联产生的数据块, 每一个数据块中包含了一批次网络交易的信息, 用于验证其信息的有效性 (防伪) 和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层。

本申请实施例可以通过数据库技术实现, 数据库 (Database), 简而言之可视为电子化的文件柜存储电子文件的处所, 用户可以对文件中的数据进行新增、查询、更新、删除等操作。所谓“数据库”是以一定方式储存在一起、能与多个用户共享、具有尽可能小的冗余度、与应用程序彼此独立的数据集合。

数据库管理系统 (Database Management System, DBMS) 是为管理数据库而设计的电脑软件系统, 一般具有存储、截取、安全保障、备份等基础功能。数据库管理系统可以依据它所支持的数据库模型来作分类, 例如关系式、XML (Extensible Markup Language, 即可扩展标记语言); 或依据所支持的计算机类型来作分类, 例如服务器群集、移动电话; 或依据所用查询语言来作分类, 例如结构化查询语言 (SQL, Structured Query Language)、XQuery; 或依据性能冲量重点来作分类, 例如最大规模、最高运行速度; 亦或其他分类方式。不论使用哪种分类方式, 一些 DBMS 能够跨类别, 例如, 同时支持多种查询语言。

在一些实施例中, 服务器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, 还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN、以及大数据和人工智能平台等基础云计算服务的云服务器。终端设备可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表、智能电视、车载终端等, 但并不局限于此。终端设备以及服务器之间可以通过有线或无线通信方式进行直接或间接地连接, 本申请实施例中不做限制。

本申请实施例, 还可以通过云技术实现, 云技术 (Cloud technology) 基于云计算商业模式应用的网络技术、信息技术、整合技术、管理平台技术、应用技术等的总称, 可以组成资源池, 按需所用, 灵活便利。云计算技术将变成重要支撑。技术网络系统的后台服务需要大量的计算、存储资源, 如视频网站、图片类网站和更多的门户网站。伴随着互联网行业的高度发展和应用, 以及搜索服务、社会网络、移动商务和开放协作等需

求的推动,将来每个物品都有可能存在自己的哈希编码识别标志,都需要传输到后台系统进行逻辑处理,不同程度级别的数据将会分开处理,各类行业数据皆需要强大的系统后盾支撑,只能通过云计算来实现。

参见图 2,图 2 是本申请实施例提供的电子设备的结构示意图,该电子设备 400 可以是图 1 中的终端设备 401,也可以是服务器(识别服务器 201、媒体服务器 202,或者二者的结合体)。该电子设备 400 包括:至少一个处理器 410、存储器 450、至少一个网络接口 420。电子设备 400 中的各个组件通过总线系统 440 耦合在一起。可理解,总线系统 440 用于实现这些组件之间的连接通信。总线系统 440 除包括数据总线之外,还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见,在图 2 中将各种总线都标为总线系统 440。

处理器 410 可以是一种集成电路芯片,具有信号的处理能力,例如通用处理器、数字信号处理器(DSP, Digital Signal Processor),或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等,其中,通用处理器可以是微处理器或者任何常规的处理器等。

存储器 450 可以是可移除的,不可移除的或其组合。示例性的硬件设备包括固态存储器,硬盘驱动器,光盘驱动器等。存储器 450 可选地包括在物理位置上远离处理器 410 的一个或多个存储设备。

存储器 450 包括易失性存储器或非易失性存储器,也可包括易失性和非易失性存储器两者。非易失性存储器可以是只读存储器(ROM, Read Only Memory),易失性存储器可以是随机存取存储器(RAM, Random Access Memory)。本申请实施例描述的存储器 450 旨在包括任意适合类型的存储器。

在一些实施例中,存储器 450 能够存储数据以支持各种操作,这些数据的示例包括程序、模块和数据结构或者其子集或超集,下面示例性说明。

操作系统 451,包括用于处理各种基本系统服务和执行硬件相关任务的系统程序,例如框架层、核心库层、驱动层等,用于实现各种基础业务以及处理基于硬件的任务。

网络通信模块 452,用于经由一个或多个(有线或无线)网络接口 420 到达其他电子设备,示例性的网络接口 420 包括:蓝牙、无线相容性认证(WiFi)、和通用串行总线(USB, Universal Serial Bus)等。

在一些实施例中,本申请实施例提供的音频数据的处理装置可以采用软件方式实现,图 2 示出了存储在存储器 450 中的音频数据的处理装置 455,其可以是程序和插件等形式的软件,包括以下软件模块:信源分离模块 4551、权重配置模块 4552、特征提取模块 4553、参数预测模块 4554,这些模块是逻辑上的,因此根据所实现的功能可以进行任意的组合或进一步拆分。将在下文中说明各个模块的功能。

在一些实施例中,终端设备或服务器可以通过运行计算机程序来实现本申请实施例提供的音频数据的处理方法。举例来说,计算机程序可以是操作系统中的原生程序或软件模块;可以是本地(Native)应用程序(APP, Application),即需要在操作系统中安装才能运行的程序,如视频 APP、音频 APP;也可以是小程序,即只需要下载到浏览器环境中就可以运行的程序。总而言之,上述计算机程序可以是任意形式的应用程序、模块或插件。

参见图 3A,图 3A 是本申请实施例提供的音频数据的处理方法的第一流程示意图,该方法可以由电子设备执行,将结合图 3A 示出的步骤进行说明。

在步骤 301 中,从音频数据提取得到至少一种信源类型分别对应的音轨数据。

示例的,音频数据文件(或者音频数据包)中分离出不同信源类型分别对应的音轨数据文件(或者,音轨数据包)。

示例的, 所述音频数据包含多个数据片段, 每个数据片段之间可以是连续的, 每个数据片段的播放时长可以是相同或者不同的。例如: 将音频数据划分为播放时长相同的多个数据片段, 或者, 将音频数据划分为播放时长不等的多个数据片段。

示例的, 音频数据可以是原生的音频数据 (例如: 有声小说、广播剧等), 也可以是从视频数据中提取的。推荐参数可以包括: 精彩程度、悲伤程度、搞笑程度、热血程度等。对应的推荐片段分别是精彩数据片段、悲伤数据片段、搞笑数据片段等。

在一些实施例中, 步骤 301 通过以下方式实现: 对音频数据进行特征提取, 得到音频数据的全局特征; 以全局特征为掩膜, 对音频数据进行信源分离, 得到音频数据中每种信源类型分别对应的音轨数据。这里, 掩膜的边界用于表征不同信源类型对应的音频数据之间的边界。

针对音频数据进行特征提取包括: 针对音频数据进行多个层次特征提取, 将每个层次得到的特征融合为全局特征。信源分离可以通过金字塔场景解析网络 (PSPN, Pyramid Scene Parsing Network) 实现, 以下分别对特征提取、信源分离进行解释说明。

在一些实施例中, 对音频数据进行特征提取处理, 得到音频数据的全局特征, 通过以下方式实现: 对音频数据进行特征提取处理, 得到音频数据的原始特征; 对原始特征进行多个层次的池化处理, 得到音频数据的多个局部特征; 对多个局部特征进行合并处理, 得到音频数据的全局特征。

示例的, 池化处理可以通过金字塔场景解析网络 (PSPN, Pyramid Scene Parsing Network) 的金字塔池化模块 (Pyramid Pooling Module) 实现, 参考图 7, 图 7 是本申请实施例提供的金字塔场景解析网络的示意图; 以下具体说明, 金字塔场景解析网络包括: 卷积神经网络 701、池化层 703 以及金字塔池化模块 (图 7 中的金字塔池化模块包括卷积层 1、卷积层 2、卷积层 3 以及卷积层 4)、上采样层 704、卷积层 706。

其中, 卷积神经网络 701 对音频数据进行特征提取, 得到音频数据的原始特征 702, 池化 (pool) 层 703 后设置的金字塔池化模块, 具体实施中可以根据提取精度设置更多的尺寸。假设金字塔共有 N 个级别, 则在每个级别后使用 1×1 卷积 (CONV), 将对应级别的通道数量降为原本的 $1/N$ 。然后通过双线性插值直接通过上采样层 704 对低维特征图进行上采样 (up sample), 得到与原始特征映射相同尺寸的特征图。金字塔场景解析网络的金字塔池化模块的每层输出不同尺寸的局部特征, 对不同级别的特征图 705 合并处理 (concat), 得到最终的全局特征。

继续参考图 7, 以全局特征为掩膜对音频数据进行信源分离的实现方式为: 将全局特征作为掩膜与金字塔场景解析网络提取到的初始层次的特征通过卷积层 706 进行卷积, 得到音频数据中每种信源类型分别对应的音轨数据对应的特征图。

示例的, 假设特征以特征矩阵的形式表征, 以全局特征为掩膜, 掩膜是与金字塔场景解析网络提取到的初始层次的特征的尺寸相同的特征矩阵, 其中全局特征对应的部分为的掩膜值为 1, 其他部分的掩膜值为 0, 将全局特征作为掩膜与金字塔场景解析网络提取到的初始层次的特征进行卷积, 能够区分不同信源类型的音频数据的频谱之间的边界, 从而在音频频谱图中表征不同信源的频谱图之间分界, 并将不同信源类型的子音频数据从整体的音频数据中单独分离出, 得到每种信源类型分别对应的音轨数据。信源类型包括: 背景音、语音。

本申请实施例中, 通过金字塔场景解析网络对音频数据进行多个层次的特征提取处理, 提升了特征提取的精确度, 进而基于提取到的全局特征与初始层次的特征提取结果进行卷积, 提升了分离不同信源类型对应的音频数据的准确性, 进而能够根据不同音源类型确定音频数据的权重值序列, 获取音频数据在音源类型方面的信息, 以提升获取推荐参数的精确度。

在步骤 302 中, 确定每个音轨数据的播放时间轴中与信源类型相关的至少一个时间段落, 并确定音频数据中每个数据片段中分别包含的时间段落。

示例的, 时间段落是音频数据的时间轴上的一段。与信源类型相关的至少一个时间段落是指信源类型对应的信源发出声音的时间段落, 信源类型对应的信源是否发出声音, 可以通过对信源类型对应音轨数据的短时能量确定。每个数据片段可以保护至少一个类型信源的时间段落, 例如: 数据片段中包含时间长度与数据片段的播放时长相同的语音的时间段落、背景音的时间段落。或者, 数据片段中包含时间长度为播放时长一半的语音的时间段落。

在一些实施例中, 参考图 3B, 图 3B 是本申请实施例提供的音频数据的处理方法的第二流程示意图, 在图 3A 中的步骤 302 通过图 3B 中的步骤 3021、步骤 3022 实现, 以下具体说明。示例的, 不限定步骤 3021 与步骤 3022 的执行顺序。

在步骤 3021 中, 当音轨数据对应的信源类型为语音时, 将音轨数据中短时能量大于能量阈值且过零率小于过零率阈值的时间段落, 作为与语音相关的时间段落。

示例的, 可以通过语音活动检测 (VAD, Voice Activity Detection) 算法获取语音相关的时间段落。短时能量, 即一帧语音信号的能量, 是帧内信号的平方和, 过零率, 即一帧语音时域信号穿过 0 (时间轴的 0 点) 的次数。语音活动检测算法的原理是, 语音数据片段的短时能量相对较大, 而过零率相对较小; 反之, 非语音数据片段的短时能量相对较小, 但是过零率相对较大。因为语音信号能量绝大部分包含在低频带内, 而噪音信号通常能量较小且含有较高频段的信息。故而可以通过测量语音信号的这两个参数并且与参数分别对应的阈值进行对比, 从而判断语音信号与非语音信号, 也即判断音轨数据中发出声音的部分和没有发出声音的部分。当音频数据的短时能量小于短时能量阈值且过零率大于过零率阈值, 则该段音频为噪音。反之, 音频数据的短时能量大于短时能量阈值且过零率小于过零率阈值时, 该段音频是语音。

在步骤 3022 中, 当音轨数据对应的信源类型为背景音时, 将音轨数据中满足筛选条件的时间段落作为与背景音相关的时间段落。

其中, 筛选条件包括以下任意一项:

条件 1、时间段落对应的响度大于响度下限值。示例的, 持续时间太短或者声音太小则有可能是杂音, 而不是背景音乐。响度下限值可以是音频数据对应的响度的中位值的预设倍数 (大于 0 且小于 1) 确定, 例如: 响度最大值与最小值的加和的平均值为响度中位值, 响度中位值的 0.5 倍为响度下限值, 将音频数据中响度小于下限值的时间段落确定为不满足筛选条件的段落。

条件 2、时间段落的长度大于长度下限值。长度下限值基于音频数据的时间长度确定, 例如: 长度下限值为音频数据的百分之一。

图 11, 图 11 是本申请实施例提供的音频数据的示意图, 其中, 音频数据 1101 的时间长度为 0 至 T6, 音频数据 1101 被划分为 6 个数据片段 (数据片段 1 至数据片段 6), 音频数据的背景音音轨 1102 中存在 T3 至 T6 的背景音信源发出声音的时间段落, 音频数据的语音音轨 1103 中存在 T1 至 T5 的语音信源发出声音的时间段落。

本申请实施例中, 用语音、背景音等类型区分音轨中每个段落, 进而能够直接定位出音频数据中的语音数据片段, 并相较于其他类型的数据片段对语音数据片段分配更高的权重值, 能够加强识别到语音数据片段的语义信息, 极大的提升了语音语义信息在精彩片段定位中的占比。

继续参考图 3A, 在步骤 303 中, 对所述音频数据中的每个数据片段基于所包含的所述时间段落长度分配对应的权重值, 并将每个所述权重值组合形成所述音频数据的权重值序列。

示例的, 为便于对音频数据进行处理, 预先将音频数据按照帧数或者时长划分为多个数据片段, 例如: 数据片段的长度为预设帧数, 或者预设时长。通过确定数据片段包含的时间段落的长度, 为数据片段分配对应的权重值。例如: 语音信源的时间段落的时间长度为 0, 不分配语音类型对应的权重值, 背景音的时间段落对应的时间长度与数据片段的播放时长相同, 对数据片段分配背景音对应的预配置权重值, 再例如: 背景音的时间段落对应的时间长度与数据片段的播放时长的一半, 则将预配置权重值的一半作为数据片段的权重值。

在一些实施例中, 当通过信源分离得到语音和背景音两种信源类型的音轨数据时, 参考图 3C, 图 3C 是本申请实施例提供的音频数据的处理方法的第三流程示意图, 图 3A 中的步骤 303 通过图 3B 中的步骤 3031 至步骤 3033 实现, 针对每个数据片段进行以下步骤 3031 至步骤 3033 处理, 以下具体说明。

示例的, 不限定步骤 3031 至步骤 3033 的执行顺序。

在步骤 3031 中, 当数据片段属于语音相关的时间段落时, 基于数据片段对应的语音的参数确定数据片段对应的权重值。

这里, 权重值与参数正相关, 参数包括以下至少之一: 语速、语调、响度。

示例的, 以影视剧视频为例进行说明, 影视剧视频的音频数据中包括语音与背景音, 语音部分通常是由演员表演的部分, 影视剧中的精彩数据片段(推荐片段)通常处于存在语音的时间段落。语音的语速、语调、响度等参数可以作为确定精彩数据片段的依据, 可以基于参数中至少一项确定数据片段对应的权重值。

在步骤 3032 中, 当数据片段属于背景音相关的时间段落时, 将预设数值作为数据片段对应的权重值。

这里, 预设数值小于任意一个语音相关的数据片段的权重值。

示例的, 影视剧视频的音频数据中包括语音与背景音, 语音部分通常是由演员表演的部分, 仅存在背景音的部分通常是影视剧视频中过场等数据片段, 可以对背景音相关时间段落分配小于语音相关数据片段的权重值。再例如: 有声小说的音频数据中, 精彩数据片段处于语音部分, 仅有背景音的时间段落可以分配更少的权重值。

在步骤 3033 中, 当数据片段不属于任意信源类型相关的时间段落时, 将零作为数据片段对应的权重值。

示例的, 当数据片段不处于任意信源类型的时间段落时, 该数据片段可能是静音或者噪声数据片段, 可以通过将数据片段的权重值置零提升获取推荐参数的准确性。

在一些实施例中, 存在以下情况: 数据片段处于任意一种信源类型的时间段落、数据片段未处于任意信源类型的时间段落、数据片段同时处于多种信源类型的时间段落(例如: 数据片段在播放时间轴中所处的时间段落, 既存在语音音轨数据也存在背景音音轨数据)。当数据片段同时处于多种信源类型的时间段落时, 获取数据片段在不同的信源类型下对应的权重值, 并每种权重值加权求和, 得到数据片段的权重值。

示例的, 为便于理解获取数据片段的权重值的过程, 以下结合附图进行说明。参考图 11, 图 11 是本申请实施例提供的音频数据的示意图, 其中, 音频数据 1101 的时间长度为 0 至 T6, 音频数据 1101 被划分为 6 个数据片段(数据片段 1 至数据片段 6), 音频数据的背景音音轨 1102 中存在 T3 至 T6 的背景音的时间段落, 音频数据的语音音轨 1103 中存在 T1 与 T2 的中点位置至 T5 的语音的时间段落, 数据片段 1 对应的时间区间为 0 至 T1, 数据片段 2 对应的时间区间为 T1 至 T2, 数据片段 3 对应的时间区间为 T2 至 T3, 数据片段 4 对应的时间区间为 T3 至 T4, 数据片段 5 对应的时间区间为 T4 至 T5, 数据片段 6 对应的时间区间为 T5 至 T6。

针对数据片段 1, 数据片段 1 与任意信源均不相关, 则数据片段 1 的权重值是 0;

针对数据片段 2 和 3, 数据片段 2 和 3 属于语音相关的时间段落, 通过上文中步骤 3031 获取权重值, 此处不再赘述, 数据片段 2 中包含的语音相关的时间段落的时长是数据偏 2 时长的一半, 因此将根据步骤 3031 获取到的权重值的一半作为数据片段 2 的权重值 q_2 , 假设数据片段 2 和 3 的权重值分别是 q_2 , q_3 。针对数据片段 4 和 5, 数据片段 4 和 5 既属于语音相关的时间段落, 也属于背景音相关的时间段落, 以数据片段 4 为例, 分别通过步骤 3031 和步骤 3032 的方式获取数据片段针对不同的信源类型的权重值, 将每个信源类型的权重值加权求和, 得到数据片段 4 的权重值 $q_4 = (aY + bB)$, 其中, Y 是数据片段 4 的语音的权重值, B 是数据片段 4 的背景音的权重值, a 和 b 分别是系数。针对数据片段 6, 数据片段 6 仅和背景音相关, 则获取背景音对应的预设值作为数据片段 6 的权重值 q_6 , 根据数据片段的先后时间顺序组合每个数据片段的权重值, 得到音频数据 1101 的权重值序列 $[0, q_2, q_3, q_4, q_5, q_6]$ 。其中, 基于语音对应的参数确定的权重值 q_3 至 q_5 相较于 0 和 q_6 更高。

本申请实施例中, 根据数据片段对应的类型选择不同的方式确定数据片段对应的权重值, 当数据片段为背景音分配预设的权重值、当数据片段属于静音或者噪声数据片段则权重值置零, 节约了获取数据片段的权重值的计算资源。当数据片段属于语音相关的时间段落, 基于语音相关的参数计算数据片段的权重值, 提升了获取语音数据片段的权重值的准确性。通过将非语音相关数据片段的权重值设置为预设值或者零, 而语音数据片段的权重值根据语音相关参数确定, 使得语音数据片段对应的权重值相较于非语音相关数据片段更高, 在视频、音频中, 推荐片段通常是存在语音的数据片段, 提升语音数据片段的权重值提升了预测每个数据片段的推荐参数的准确性。

在一些实施例中, 当通过信源分离仅得到背景音一种信源类型的音轨数据时, 参考图 3D, 图 3D 是本申请实施例提供的音频数据的处理方法的第四流程示意图, 图 3A 中的步骤 303 通过图 3D 中的步骤 3034 至步骤 3035 实现, 针对每个数据片段进行以下步骤 3034 和步骤 3035 的处理, 以下具体说明。

示例的, 不限定步骤 3034 与步骤 3035 的执行顺序。

在步骤 3034 中, 当数据片段包含的时间段落属于背景音相关的时间段落时, 基于数据片段对应的背景音的参数确定数据片段对应的权重值。

这里, 权重值与参数正相关, 参数包括以下至少之一: 响度、音调。

示例的, 假设音频数据是音乐会的音频数据, 则仅包括背景音信源而不一定存在语音, 语调、响度等参数可以作为确定精彩数据片段的依据, 可以基于参数中至少一项确定数据片段对应的权重值。

在步骤 3035 中, 当数据片段包含的时间段落不属于任意信源类型相关的时间段落时, 将零作为数据片段对应的权重值。

示例的, 步骤 3035 与步骤 3033 的内容相同, 此处不再赘述。

本申请实施例中, 当多媒体数据中不存在语音时, 针对属于背景音的数据片段分配预设的权重值、针对属于静音或者噪声的数据片段则权重值置零, 节约了获取数据片段的权重值的计算资源。

继续参考图 3A, 在步骤 304 中, 从每个数据片段提取音频特征, 将每个数据片段的与音频特征组合形成音频数据的音频特征序列, 并对音频特征序列进行编码, 得到音频数据的注意力参数序列。

示例的, 从每个数据片段中提取音频特征通过以下方式实现: 对音频数据进行特征提取, 得到单独的频域特征或者单独的时域特征。

在一些实施例中, 在步骤 304 之前, 可以通过以下方式获取音频特征, 针对音频数据中每个数据片段进行以下处理: 提取数据片段的时域信号特征与频域信号特征; 基于

每个音轨数据的播放时间轴中与信源类型相关的至少一个时间段落,确定时域信号特征对应的一维时域权重值,以及确定频域信号特征对应的二维频域权重值;对一维时域权重值与时域信号特征的乘积进行多个层次的卷积,得到时域音频特征;对二维频域权重值与频域信号特征的乘积进行多个层次的卷积,得到频域音频特征;对时域音频特征进行缩放,得到二维时域音频特征;对二维时域音频特征与频域音频特征进行融合处理,得到数据片段的音频特征。

示例的,时域音频特征为一维特征,可以通过对时域音频特征进行缩放,便于将时域特征与频域特征进行融合,例如:通过 reshape 函数对一维特征进行处理,得到不改变元素数量的二维特征。

参考图 6B,图 6B 是本申请实施例提供的音频处理模型的第二示意图;步骤 304 可以通过图 6B 中的音频语义信息提取模块 605 实现,音频语义信息提取模块 605 的结构为双流型,包括时域支路 606 以及频域支路 607,音频数据的时域信息、权重值序列输入时域支路 606,时域支路 606 包括多个一维卷积层(一维卷积层 1、.....一维卷积层 n),音频数据的频域信息、权重值序列输入频域支路 607,频域支路 607 包括多个二维卷积层(二维卷积层 1、.....二维卷积层 n)。特征融合层 608 用于融合两条支路中各个层次的卷积层输出的频域特征或者时域特征。

以下具体说明,参考图 8,图 8 是本申请实施例提供的音频语义信息提取模块的示意图,图 8 是图 6B 中音频语义信息提取模块 605 的细化结构图;音频语义信息提取模块的输入为视频的原始音频数据(表征为音频采样点序列)。音频数据被划分为多个数据片段(例如:按照以下方式划分:每个数据片段包括至少一帧,或者每个数据片段的长度相等)。将基于音频数据生成基础特征图(logmel)作为频域信息,并输入到频域支路 607,音频数据的音频采样点序列(时域信息)被输入时域支路 606。权重分配单元 604 输出的权重值序列通过全连接层 801、全连接层 802 的处理,分别生成与时域信号特征和频域信号特征相同维度的时间权重向量,然后分别与时域信号特征和频域信号特征进行对应位置相乘。

时域支路 606 包括大量的一维卷积层(一维卷积层 803、一维卷积层 804、一维卷积层 806、一维卷积层 808)以及一维最大池化层(一维最大池化层 805、一维最大池化层 807、一维最大池化层 809),在时域信号特征中使用大量的卷积层能够直接学习到音频数据的时域特性,尤其是像音频响度和采样点幅度的信息。经过大量的一维卷积层后,把生成的一维序列通过变形层 810 缩放(resize)成为一个二维图谱(wave graph)形式的特征图,这种处理使得时域支路与频域支路输出的特征的尺寸相同,便于进行融合处理。

示例的,在时域支路的一维卷积的过程中,中间结果通过变形层(变形层 811、变形层 812)缩放为二维图谱(wave graph),通过合并层(例如:合并层 813、合并层 815)、二维卷积层(例如:二维卷积层 814、二维卷积层 816)与频域支路 607 的中间结果进行多个层次的合并,使得最终得到的音频特征能够融合不同尺寸、层次的频域特征与时域特征。

频域支路 607 输出的频域信息可以为采用梅尔频域的 log-mel 频谱,频域支路 607 包括大量的二维卷积层(二维卷积层 821、二维卷积层 823、二维卷积层 825)以及二维最大池化层(二维最大池化层 822、二维最大池化层 824),在频域信号特征中使用大量的卷积层能够直接学习到音频数据的频域特性。经过大量的二维卷积层后,得到二维特征图,二维特征图的维度与时域支路 606 输出的特征图的维度相同。

示例的,在频域支路进行二维卷积的过程中,中间结果通过合并层(例如:合并层 813、合并层 815)、二维卷积层(例如:二维卷积层 814、二维卷积层 816)与时域支路

606 的中间结果进行多个层次的合并,使得最终得到的音频特征能够融合不同尺寸、层次的频域特征与频域特征。

示例的,变形层可以通过 reshape 函数(将指定的矩阵变换成特定维数矩阵一种函数,且矩阵中元素个数不变,函数可以重新调整矩阵的行数、列数、维数。)对特征图进行变形。

在一些实施例中,对二维时域音频特征与频域音频特征进行融合处理,得到数据片段的音频特征,通过以下方式实现:对二维时域音频特征与频域音频特征进行叠加处理,对叠加处理得到的叠加特征进行二维卷积,得到二维卷积结果,获取二维卷积结果的最大叠加特征(Max)与平均叠加特征(Mean);对最大叠加特征与平均叠加特征之间的加和进行线性激活,得到数据片段的音频特征。

作为叠加处理的示例,二维时域音频特征与频域音频特征可以分别表征为特征矩阵,对二维时域音频特征的特征矩阵、频域音频特征的特征矩阵进行线性相加,得到叠加特征,采用特征矩阵的形式表征叠加特征。

示例的,本申请实施例中音频特征以向量形式表征。线性激活,也即通过 Relu 函数对最大叠加特征与平均叠加特征之间的加和进行激活处理,得到数据片段的音频特征。继续参考图 8,特征融合模块 608 中的合并层 817 合并两个支路输出的特征图,合并处理使得时域和频域保持信息上的互补,同时还能够让高层网络感知到底层网络信息。合并层 817 输出每个数据片段的二维频域特征图,将二维频域特征图输入到二维卷积神经网络层 818 中。得到二维卷积结果,基于二维卷积结果获取二维卷积神经网络层最后一维度特征的平均值(mean)和最大值(max),将求得的平均值与最大值通过合并层 819 进行相加,加和通过激活层 820 利用线性激活函数(relu),生成最终的音频语义特征向量(音频特征)。将每个数据片段的音频语义特征向量根据数据片段对应的时间顺序组合,得到音频特征序列。

本申请实施中,将时域特征转换为与频域特征相同的维度的特征,降低了对音频数据的时域特征、频域特征进行融合的复杂度,节约了计算资源,提升了特征融合的准确性,通过频域特征与时域特征相互融合,能够从不同方面获取音频中所包含的信息,使得音频特征表征的信息量更全面,提升了获取音频特征的精确度。

在一些实施例中,也可以仅采集音频数据的频域特征或者时域特征作为音频的音频特征,通过仅采集一种域的特征的方式,可以提升计算速度,节约计算资源。

在一些实施中,参考图 3E,图 3E 是本申请实施例提供的音频数据的处理方法的第五流程示意图,图 3A 中的步骤 304 中的对音频特征序列进行编码,得到音频数据的注意力参数序列通过图 3E 中的步骤 3041 至步骤 3043 实现。

在步骤 3041 中,针对音频特征序列中每个音频特征执行以下处理:基于注意力机制,将音频特征与其他数据片段的每个音频特征分别融合得到音频特征对应的每个加权相关度。

这里,其他数据片段是音频数据中除当前获取加权相关度的数据片段之外的数据片段。

示例的,以下以音频特征 A 进行举例,融合处理是将音频特征 A 的嵌入向量与任意一个其他数据片段的音频特征的嵌入向量进行内积,并将内积结果与音频特征 A 相乘,得到音频特征 A 的一个加权相关度,针对每个其他数据片段的音频特征获取加权相关度,则得到音频特征 A 对应的每个加权相关度。

在一些实施例中,在步骤 3041 之前,通过以下方式确定每个音频特征的嵌入向量:通过全连接层对音频数据的每个数据片段的音频特征进行全连接,得到每个音频特征的嵌入向量。

示例的，参考图 9，图 9 是本申请实施例提供的注意力模块中编码的原理示意图。假设：音频特征序列包括 a_1 至 a_n 等多个音频特征，将每个两个数据片段对应的音频特征通过全连接层进行处理，得到音频特征对应的一维嵌入（embedding）向量（两个向量的阶数相同）。

5 在一些实施例中，步骤 3041 通过以下方式实现：针对音频特征与其他数据片段的每个音频特征执行以下处理：对音频特征的嵌入向量与其他数据片段的嵌入向量相乘，得到音频特征与其他数据片段的音频特征之间的相关度；将音频特征与相关度相乘，得到音频特征对应的加权相关度。

示例的，继续参考图 9，音频特征表征为一维嵌入向量形式，对两个一维嵌入向量进行内积计算，得到相关度 m ，例如：音频特征序列中的第 1 个音频特征表征为 a_1 与第 i 个音频特征表征为 a_i ，音频特征 a_1 与音频特征 a_i 相乘得到的相关度为 m_{1i} 。将相关度与音频特征 a_1 进行相乘，得到加权相关度 c_{1i} 。

在步骤 3042 中，将每个加权相关度相加得到音频特征对应的注意力参数。

示例的，继续参考图 9，音频特征 a_1 与音频特征 a_i 之间的加权相关度 c_{1i} ，
15 $m_{1i} \times a_1 = c_{1i}$ 。将同一个音频特征对应的每个加权相关度相加，可以得到该音频特征对应的注意力参数 W 。例如：音频特征 a_1 的注意力参数 $W_1 = \sum_{i=1}^n c_{1i}$ 。

在步骤 3043 中，基于每个音频特征对应的数据片段的顺序，将每个注意力参数组合形成音频数据的注意力参数序列。

示例的，数据片段的顺序是指数据片段在音频数据中的时间先后顺序，注意力参数序列中每个注意力参数与每个数据片段一一对应，每个注意力参数是根据注意力参数对应的数据片段的时间先后顺序组合为注意力参数序列的，权重值序列中每个权重值也与每个数据片段一一对应，每个权重值是根据权重值对应的数据片段的时间先后顺序组合为权重值序列的。

本申请实施例，通过基于注意力机制对音频特征进行融合获取注意力参数，基于注意力参数能够更准确地确定推荐参数，进而提升了确定推荐片段的准确性，解决了缺乏播放记录数据的音频数据或者视频数据难以确定推荐片段的问题。

继续参考图 3A，在步骤 305 中，将注意力参数序列与权重值序列融合得到每个数据片段的融合参数，并基于每个融合参数确定每个数据片段的推荐参数。

示例的，融合处理是将注意力参数序列与权重值序列进行相乘，注意力参数序列与权重值序列中所包含的元素的数量是相同的。

在一些实施例中，步骤 305 通过以下方式实现：针对每个数据片段执行以下处理：从注意力参数序列获取数据片段对应的注意力参数，将数据片段的权重值与数据片段的注意力参数相乘，得到数据片段的融合参数；对融合参数进行归一化处理，得到数据片段的推荐参数。

示例的，假设音频数据的权重值序列是 $[Q1, Q2, \dots, Qn]$ ，注意力参数序列是 $[Z1, Z2, \dots, Zn]$ ，则音频数据中的第一个数据片段的融合参数是 $Q1 \times Z1$ ，也即，第一个数据片段的权重值与注意力参数的乘积。

示例的，归一化处理是通过 softmax 函数进行置信度预测。以推荐参数是精彩程度进行举例，针对影视剧视频，视频中存在语音的部分为精彩数据片段的概率更高，基于语音信源的音轨数据分配对应的权重值，语音信源的权重值高于背景音部分的权重值，使得语音数据片段对应的精彩程度置信度相较于背景音数据片段对应的精彩程度置信度更高。

本申请实施例中，通过将注意力参数与权重值进行融合，基于融合参数确定推荐参数，将频域时域的信息、信源类型的信息结合，使得推荐参数能够更全面地量化表征音

频数据的信息，提升了确定推荐参数的精确度。

在步骤 306 中，基于每个数据片段的推荐参数，确定音频数据中的推荐片段。

在一些实施例中，通过以下任意一种方式确定音频数据的推荐片段：

1、基于每个数据片段的推荐参数对每个数据片段进行降序排序，将降序排序的头部的至少一个数据片段作为音频数据的推荐片段。例如：对每个数据片段的精彩程度进行降序排序，将头部的预设数量的数据片段作为精彩数据片段，预设数量与音频数据的数据片段总数正相关，例如：预设数量为数据片段总数的百分之一。

2、将推荐参数大于推荐参数阈值的数据片段作为推荐片段。示例的，推荐参数阈值可以是每个数据片段的推荐参数的中位值，或者中位值的预设倍数（例如：1.5 倍， $1 < \text{预设倍数} < 2$ ），假设，最大的推荐参数为 0.9，最小的推荐参数为 0，取中位值 0.45 为推荐参数阈值，将精彩程度大于 0.45 的数据片段作为精彩数据片段。再假设，最大的推荐参数为 0.9，最小的推荐参数为 0，取中位值的 1.1 倍为推荐参数阈值，则推荐参数阈值为 0.495。

本申请实施例中，通过推荐参数量化音频数据中每个数据片段与信源之间的相关程度，通过推荐参数表征音频数据属于某个特定类型的推荐片段的概率，选取推荐参数最高多个数据片段作为推荐片段，选取得到的推荐片段可以表征音频数据中的特定类型的位置，相较于单纯从频域、时域的层面来预测，结合了不同信源识别更加全面，从而基于每个数据片段的推荐参数可以准确识别出有价值的推荐片段，为用户提供准确的参考信息。

在一些实施例中，在步骤 305 之后，还可以基于每个数据片段的推荐参数，生成音频数据的推荐参数曲线；响应于播放触发操作，在播放界面显示音频数据的推荐参数曲线。

这里，推荐参数曲线的横坐标为音频数据的播放时间，推荐参数曲线的纵坐标为推荐参数。

示例的，推荐参数曲线的横坐标与音频数据的播放时间一一对应，推荐参数曲线的纵坐标越高，则推荐参数越大。参考图 10A，图 10A 是本申请实施例提供的播放界面的第一示意图。播放界面 101A 为视频播放器的播放界面，推荐参数为精彩程度，精彩程度曲线 106A 显示在不遮挡视频画面的区域，精彩数据片段 107A 被标注出。进度条 105A 中的滑块 103A 所在位置是视频当前播放的时刻对应的位置。进度条 105A 可以表征播放时间。精彩程度曲线 106A 的高低可以表征精彩程度的大小。

示例的，播放触发操作可以是针对音频或者视频的。播放界面可以是音频播放界面或者视频播放界面，则音频播放界面，播放音频数据（对应音频播放场景，音频数据），视频播放界面，对应视频播放场景，音频数据是从视频数据提取的。

在一些实施例中，在步骤 306 之后，还可以在播放界面显示推荐片段的标签，其中，标签用于表征推荐片段的时间段落；响应于针对任意一个标签的选择操作，跳转到选中的标签对应的推荐片段的起点开始播放。

示例的，选择操作可以是点击操作，或者将进度条滑块拖动到标签的操作，参考图 10B，图 10B 是本申请实施例提供的播放界面的第二示意图。滑块 103A 被拖动到标签 104A 的位置，视频画面切换为精彩数据片段 107A 的起点位置的画面。

在一些实施例中，本申请实施例提供的音频数据的处理方法通过音频处理模型实现。信源分离通过调用音频处理模型的金字塔场景解析网络实现，从每个数据片段提取音频特征通过调用音频处理模型的音频语义信息提取模块实现，编码与融合处理通过调用音频处理模型的注意力模块实现。

参考图 6A，图 6A 是本申请实施例提供的音频处理模型的第一示意图。音频处理模

型包括金字塔场景解析网络 601、权重配置模块 610、音频语义信息提取模块 605 以及注意力模块 609。金字塔场景解析网络 601 用于执行步骤 301，权重配置模块 610 用于执行步骤 303，音频语义信息提取模块 605 用于执行步骤 304，注意力模块 609 用于执行步骤 609。

5 音频数据输入金字塔场景解析网络 601，金字塔场景解析网络 601 对音频数据进行信源分离到至少一种信源类型对应的音轨数据，权重配置模块 610 用于实现上文中的步骤 303，权重配置模块 610 确定音轨数据中与信源关联的时间段落，并对时间段落分配对应的权重值，将权重值输出到音频语义信息提取模块 605、注意力模块 609。音频数据输入到音频语义信息提取模块 605（音频语义信息提取模块的具体结构参考上文中图 10 6B 以及图 8），音频语义信息提取模块 605 对音频数据进行时域、频域两方面的特征提取处理，并将融合时域、频域信息的音频特征序列输出到注意力模块 609，注意力模块 609 是用于运行注意力机制的算法模块，注意力模块 609 通过注意力机制基于权重值序列与音频特征序列进行参数预测，得到推荐参数，制作推荐参数曲线。

其中，通过以下方式训练音频处理模型：基于音频数据的每个实际推荐片段的标签值（标签值也即实际推荐片段的推荐参数，正样本的标签值为 1），组合形成音频数据的实际推荐参数序列；基于音频数据的每个数据片段的推荐参数，组合形成音频数据的预测推荐参数序列；基于实际推荐参数序列与预测推荐参数序列获取音频处理模型的交叉熵损失；将交叉熵损失除以音频数据的数据片段数量，得到平均预测损失，基于平均预测损失对音频处理模型进行反向传播，得到更新后的音频处理模型。

20 示例的，训练数据存在人工标注的标签值，标签值能够表征实际上哪些数据片段为推荐片段（精彩数据片段）的概率，其中，推荐片段被标注为 1（正样本），非推荐片段被标注为 0（负样本），在进行损失函数计算时，一个视频对应的所有的标签值可以组成一个实际推荐参数序列（由 0、1 组成的序列）。例如：视频划分为 N 个数据片段，N 是正整数，推荐片段是精彩片段，人工标注出视频中的精彩片段，根据每个数据片段在 25 视频中的时间从前至后的顺序组合标签值为实际推荐参数序列，实际推荐参数序列表征为 [1、0、1.....0]。

在一些实施例，当音频数据为视频中截取的音频数据时，可以在音频特征的基础上结合图像信息确定精彩数据片段。可以通过以下方式实现：对视频的图像数据进行图像特征提取，将图像特征与对应的音频特征进行融合，得到融合的视频特征，基于视频 30 特征执行注意力机制，得到注意力参数序列，基于注意力参数序列与权重值序列确定推荐参数序列。

在一些实施例，当音频数据为视频中截取的音频数据时，可以基于视频的图像特征识别到的推荐片段优化基于音频数据识别到的推荐片段，通过以下方式实现：对视频的图像数据进行图像识别，基于识别得到的包括人物的视频图像，确定视频中包括人物 35 的数据片段时间。将推荐参数大于推荐参数阈值，且对应的视频数据片段中包括人物的视频数据片段作为推荐片段。

示例的，还可以通过以下方式确定视频数据的精彩数据片段：对视频的图像数据（视频画面）进行特征提取处理，得到视频的图像语义特征序列；对视频的图像数据进行图像识别，得到视频中包括人物的数据片段时间，并基于人物数据片段时间对视频分配对应的权重值序列。基于图像语义特征序列获取注意力参数，得到注意力参数序列，基于 40 图像数据的注意力参数序列与权重序列得到视频画面对应的推荐参数。对视频画面的推荐参数与音频数据的推荐参数加权求和，得到加权推荐参数，将加权推荐参数大于加权推荐参数阈值的视频数据片段作为推荐片段。

本申请实施例从音频侧的角度来对整个视频进行多个域内以及多层信息的分析，能

够快速的定位出整个音频中的推荐片段（例如：精彩数据片段、热血数据片段、悲伤数据片段或者搞笑数据片段等），从而基于音频的推荐片段能够判断出视频中的推荐片段的时间段落在时间轴中的位置。从而在不依赖音频数据的播放记录数据的情况下，就可以准确识别推荐片段，为用户提供准确的参考信息，提升了用户体验。能够为播放器提供视频推荐参数曲线，以供观众将播放进度条由当前播放位置跳转到推荐片段的位置，提升观众对播放器的使用体验。

下面，将说明本申请实施例在一个实际的应用场景中的示例性应用，本申请实施例提供的音频数据的处理方法可以应用在如下应用场景中：

1、在不同平台端（pc\tv\android\ios）播放长视频的过程中，在播放器中能够显示视频的时间轴进度条关联的热度信息。热度信息通常是基于视频的播放记录数据（播放量、点击量、弹幕或者评论数量等）计算得到的，但是针对于新上映的电影或者影视剧的视频，这些视频没有播放记录数据。或者，针对于小众视频没有足够播放记录数据确定热度。本申请实施例提供的音频数据的处理方法可以生成推荐参数曲线来替代热度信息，推荐参数可以是精彩程度，向用户展示视频中的精彩数据片段与精彩程度曲线，用户根据精彩程度曲线或者精彩数据片段标签可直接跳转到精彩数据片段进行观看或收听，提升用户的观看体验。

2、针对于某些短视频平台中的影视剧二创短视频制作，用户往往是先自己观看影视剧之后再从整个剧集中定位精彩数据片段，得到精彩数据片段锦集。基于定位得到的精彩数据片段锦集，进行二创短视频集锦类制作。本申请实施例提供的音频数据的处理方法可以为二创用户提供精彩程度曲线，用户可以根据曲线一目了然地确定视频中的精彩数据片段，一键定位、截取整个视频中的精彩数据片段的画面，然后二创用户可以直接根据截取的结果来进行接下来的短视频生成工作，大幅度提升效率，避免了人工分辨精彩数据片段而浪费时间。

下面，以推荐参数为精彩程度，以音频数据为影视剧的视频的音频数据为例进行说明。参考图 5，图 5 是本申请实施例提供的音频数据的处理方法的一个可选的流程示意图，下面以电子设备为执行主体，将结合图 5 的步骤进行说明。

步骤 501 中，获取待处理的视频文件。

示例的，待处理的视频文件可以是影视剧或者电影的视频文件。视频文件由视频画面帧与音频数据组成，音频数据中可以提取到至少一种信源类型对应的音轨数据。参考图 4A，图 4A 是本申请实施例提供的视频中提取的音频数据的示意图；图 4A 中由上至下，分别是视频画面帧的示意图（表征视频的预览画面）、音频数据的音频特征图、音轨数据的音频采样序列图以及推荐参数曲线的示意图。推荐参数曲线的横坐标表示时间，纵坐标表示推荐参数。

步骤 502 中，基于视频文件的音频数据调用音频处理模型进行精彩置信度预测处理，得到音频数据的精彩置信度曲线以及精彩数据片段。

参考图 6A，图 6A 是本申请实施例提供的音频处理模型的第一示意图。音频处理模型包括金字塔场景解析网络 601、权重配置模块 610、音频语义信息提取模块 605 以及注意力模块 609。音频数据输入金字塔场景解析网络 601，金字塔场景解析网络 601 对音频数据进行信源分离到至少一种信源类型对应的音轨数据，权重配置模块 610 确定音轨数据中与信源关联的时间段落，并对时间段落分配对应的权重值，将权重值输出到音频语义信息提取模块 605、注意力模块 609。音频数据输入到音频语义信息提取模块 605，音频语义信息提取模块 605 对音频数据进行时域、频域两方面的特征提取处理，并将融合时域、频域信息的音频特征序列输出到注意力模块 609，注意力模块基于权重值序列与音频特征序列进行参数预测处理，得到推荐参数，制作推荐参数曲线。

以下对音频处理模型中各模块进行解释说明,参考图 6B,图 6B 是本申请实施例提供的音频处理模型的第二示意图。金字塔场景解析网络 601、权重配置模块 610 中的语音定位单元 603 进行对整条音轨中的语音段落进行毫秒级别的定位。语音定位单元 603 采用语音活动检测算法,金字塔场景解析网络 601 为金字塔场景解析网络 (PSPN, Pyramid Scene Parsing Network),通过金字塔形式的卷积层网络,由大到小的感受域能够更好的对分离细节进行识别定位。使用金字塔场景解析网络能够更精准的将音频频谱图中不同的特征进行分离,尤其是在金字塔卷积层中的小卷积层,能够学习到在音频频谱图中不同信源的频谱图之间分界的边缘性,以不同信源的特征的边缘为掩膜对频谱图进行分离处理,使得分离得到的不同信源的音轨数据更准确。视频的原始音轨被输入金字塔场景解析网络 601,输出为分离的背景音音轨和语音音轨等音轨数据(图 6B 中的音轨数据 602)。然后采用语音活动检测开源算法进行对语音音轨中的语音段落进行定位,从而得到整个音轨中的语音的时间段落。

示例的,金字塔场景解析网络 601 基于金字塔场景解析网络搭建的信源分离模型对整个视频的音轨进行分离,将音轨中的语音信息和背景音信息分别进行分裂,单独存储成为音轨数据(音轨文件)。语音定位单元 603 基于语音活动检测算法对语音的音轨数据中的语音数据片段进行定位,得到存在语音的时间段落,权重分配单元 604 对每个语音的时间段落的权重进行设置。语音的时间段落被分配的权重值相较于纯背景音的时间段落更高。

本申请实施例中,在金字塔场景解析网络中,金字塔池化模块(Pyramid Pooling Module)生成的不同层次的特征图最终被合并层(concat)合并,并将合并得到的特征图拼接起来,再输入到全连接层以进行分类。金字塔场景解析网络通过金字塔池化模块的多个层次的卷积层输出包含不同尺度、不同子区域间的局部信息,并在金字塔场景解析网络的最终的卷积层特征图上构造全局先验信息。该全局先验信息旨在消除卷积神经网络对图像分类输入大小固定的限制。

参考图 7,图 7 是本申请实施例提供的金字塔场景解析网络的示意图;以下具体说明,图 7 的是金字塔场景解析网络图 6A 以及图 6B 中金字塔场景解析网络 601 的细化结构示意图,卷积神经网络 701 对音频数据进行特征提取,得到音频数据的原始特征 702,池化(pool)层 703 后设置的金字塔模块(包括卷积层 1、卷积层 2、卷积层 3 以及卷积层 4,具体实施中可以根据提取精度设置更多的尺寸)可以融合四种不同金字塔尺度的特征:卷积层 1 突出显示的是最粗糙级别的单个全局池化输出,金字塔模块的多个层次不同尺寸的卷积层将原始特征映射划分为不同的子区域,并形成针对不同位置的局部特征。金字塔模块中不同层次的卷积层输出不同尺寸的局部特征。为了维护全局特性的权重,假设金字塔共有 N 个级别,则在每个级别后使用 1×1 卷积(CONV),将对应级别的通道数量降为原本的 $1/N$ 。然后通过双线性插值直接通过上采样层 704 对低维特征图进行上采样(up sample),得到与原始特征映射相同尺寸的特征图。最后,将金字塔模块输出的不同级别的特征图 705 合并处理(concat),通过卷积层 706 对合并处理的结果进行卷积,得到最终的金字塔全局特征。继续参考图 7,可以看出金字塔场景解析模型的架构呈一个金字塔形状。该模型输入图像后,使用预训练的带空洞卷积层提取特征图,空洞卷积(Atrous Convolutions)又称为扩张卷积(Dilated Convolutions),在卷积层中引入了扩张率(dilation rate),扩张率定义了卷积核处理数据时各数据值的间距。由于引入池化层会导致全局信息的损失,空洞卷积层的作用是在不使用池化层的情况下提供更大的感受野。最终的特征映射大小是输入图像的 $1/8$,然后将该特征输入到金字塔池化模块中,模型使用金字塔场景解析网络中金字塔池化模块来收集上下文信息。金字塔池化模块为 4 层金字塔结构,池化内核覆盖了图像的全部、一半和小部分。它们被融合

为全局先验信息（全局特征），在最后部分将之前的全局特征映射与原始特征映射合并起来再进行卷积（以全局特征为掩膜，分离原始特征中的语音与背景音），生成语音、背景音的最终分割特征图。

参考图 4B，图 4B 是本申请实施例提供的音轨数据示意图；图 4B 中上图为音轨波形图（采样序列图），下图为语音对应的音轨特征图，音轨特征图中空白部分为舍弃的噪音部分。示例的，通过金字塔场景解析网络搭建的信源分离模型可以分离出原始音轨中的语音、背景音分别对应的音轨数据。基于音轨数据可以使用语音活动检测算法（例如：WebRTC 语音活动检测算法）对具体的音频冲激信号段落进行定位。语音活动检测算法，是基于短时能量（STE, Short Time Energy）和过零率（ZCC, Zero Cross Counter）确定音频是否为语音的算法。短时能量，即一帧语音信号的能量，是帧内信号的平方和，过零率，即一帧语音时域信号穿过 0（时间轴）的次数。语音活动检测算法的原理是，语音数据片段的短时能量相对较大，而过零率相对较小；反之，非语音数据片段的短时能量相对较小，但是过零率相对较大。因为语音信号能量绝大部分包含在低频带内，而噪音信号通常能量较小且含有较高频段的信息。故而可以通过测量语音信号的这两个参数并且与参数分别对应的阈值进行对比，从而判断语音信号与非语音信号。当音频数据的短时能量小于短时能量阈值且过零率大于过零率阈值，则该段音频为噪音。反之，音频数据的短时能量大于短时能量阈值且过零率小于过零率阈值时，该段音频是语音。参考图 4C，图 4C 是本申请实施例提供的时间段落示意图；框 401C 圈出的时间段落为语音的时间段落，同理地，图 4C 中圈出的每个框对应的波形均为语音的时间段落。

继续参考图 6B，音频语义信息提取模块 605 的结构为双流型，包括时域支路 606 以及频域支路 607，音频数据的时域信息、权重值序列输入时域支路 606，时域支路 606 包括多个一维卷积层（一维卷积层 1、.....一维卷积层 n），音频数据的频域信息、权重值序列输入频域支路 607，频域支路 607 包括多个二维卷积层（二维卷积层 1、.....二维卷积层 n）。特征融合层 608 用于融合两条支路中各个层次的卷积层输出的频域特征或者时域特征。

以下具体说明，参考图 8，图 8 是本申请实施例提供的音频语义信息提取模块的示意图；音频语义信息提取模块的输入为视频的原始音频数据（表征为音频采样点序列）。音频数据被划分为多个数据片段（例如：通过以下方式划分：每个数据片段包括至少一帧，或者每个数据片段的长度相等）。将基于音频数据生成基础特征图（logmel）作为频域信息，并输入到频域支路 607，音频数据的音频采样点序列（时域信息）被输入时域支路 606。权重分配单元 604 输出的权重值序列通过全连接层 801、全连接层 802 的处理，分别生成与时域信号特征和频域信号特征相同维度的时间权重向量，然后分别与时域信号特征和频域信号特征进行对应位置相乘。

时域支路 606 包括大量的一维卷积层（一维卷积层 803、一维卷积层 804、一维卷积层 806、一维卷积层 808）以及一维最大池化层（一维最大池化层 805、一维最大池化层 807、一维最大池化层 809），在时域信号特征中使用大量的卷积层能够直接学习到音频数据的时域特性，包括音频响度和采样点幅度的信息。经过大量的一维卷积层后，把生成的一维序列通过变形层 810 缩放（resize）成为一个二维图谱（wave graph）形式的特征图，这种处理使得时域支路与频域支路输出的特征的尺寸相同，便于进行融合处理。

示例的，在时域支路进行一维卷积的过程中，中间结果通过变形层（变形层 811、变形层 812）缩放为二维图谱（wave graph），通过合并层（例如：合并层 813、合并层 815）、二维卷积层（例如：二维卷积层 814、二维卷积层 816）与频域支路 607 的中间结果进行多个层次的合并，使得最终得到的音频特征能够融合不同尺寸、层次的频域特征与时域特征。

频域支路 607 输出的频域信息可以为采用梅尔频域的 logmel 频谱, 频域支路 607 包括大量的二维卷积层 (二维卷积层 821、二维卷积层 823、二维卷积层 825) 以及二维最大池化层 (二维最大池化层 822、二维最大池化层 824), 在频域信号特征中使用大量的卷积层能够直接学习到音频数据的频域特性。经过大量的二维卷积层后, 得到二维特征图, 二维特征图的维度与时域支路 606 输出的特征图的维度相同。

示例的, 在频域支路进行二维卷积的过程中, 中间结果通过合并层 (例如: 合并层 813、合并层 815)、二维卷积层 (例如: 二维卷积层 814、二维卷积层 816) 与时域支路 606 的中间结果进行多个层次的合并, 使得最终得到的音频特征能够融合不同尺寸、层次的频域特征与频域特征。

示例的, 变形层可以通过 reshape 函数 (将指定的矩阵变换成特定维数矩阵一种函数, 且矩阵中元素个数不变, 函数可以重新调整矩阵的行数、列数、维数。) 对特征图进行变形。

特征融合模块 608 中的合并层 817 合并两个支路输出的特征图, 合并处理使得时域和频域保持信息上的互补, 同时还能够让高层网络感知到底层网络信息。合并层 817 输出每个数据片段的二维频域特征图, 将二维频域特征图输入到二维卷积神经网络层 818 中。得到二维卷积结果, 确定二维卷积结果的平均值 (Mean) 和最大值 (Max), 将求得的平均值与最大值通过合并层 819 进行相加, 加和通过激活层 820 利用线性激活函数 (Relu), 生成最终的音频语义特征向量 (音频特征)。将每个数据片段的音频语义特征向量组合, 得到音频特征序列。

继续参考图 6B, 注意力模块 609 接收权重值序列与音频特征序列, 注意力模块基于音频特征序列编码得到注意力参数序列, 基于注意力参数序列与权重值序列预测每个数据片段的推荐参数。参考图 9, 图 9 是本申请实施例提供的注意力模块中编码的原理示意图。

示例的, 假设音频特征序列包括 a_1 至 a_n 等多个音频特征, 将每个两个数据片段对应的音频特征通过全连接层进行处理, 得到音频特征对应的一维嵌入 (embedding) 向量 (两个向量的阶数相同), 对两个一维嵌入向量进行内积的计算, 得到相关度 m , 例如: 音频特征 a_1 与音频特征 a_i 之间的相关度为 m_{1i} 。将相关度与音频特征对应的向量进行相乘, 得到加权相关度信息量 c (上文的加权相关度)。再例如: 音频特征 a_1 与音频特征 a_i 之间的加权相关度信息量 c_{1i} , $m_{1i} \times a_1 = c_{1i}$ 。将音频特征对应的每个加权相关度信息量相加, 可以得到音频特征对应的注意力参数 W 。例如: 音频特征 a_1 的注意力参数 $W_1 = \sum_{i=1}^n c_{1i}$ 。

通过上述方式获取所有数据片段的音频特征对应的注意力参数, 针对每个数据片段, 将该数据片段对应的注意力参数 W 与该数据片段对应的权重值 L 进行相乘, 得到最终的输出的特征序列 Q (特征序列 Q 的粒度可以为帧级别), 通过二分类层对每个粒度的特征节点进行归一化处理: 二分类的标签为 1-0, 1 类别的后验概率为该特征节点的置信度 (精彩程度), 也即, 代表该特征节点的特征为精彩的概率; 针对整个推荐参数序列执行归一化处理 (例如通过 softmax 函数), 即可得到精彩程度曲线。可以设置对应的精彩程度阈值, 将精彩程度大于精彩程度阈值的数据片段作为精彩数据片段, 小于精彩程度阈值的数据片段作为非精彩数据片段。

在一些实施例中, 在训练过程中训练数据存在人工标注的标签 (label), 标签能够表征实际上哪些数据片段为推荐片段 (精彩数据片段), 其中, 推荐片段被标注为 1 (正样本), 非推荐片段被标注为 0 (负样本), 在进行损失函数计算时, 一个视频对应的所有的标签可以组成一个 0-1 序列。基于推荐参数序列与标签序列计算交叉熵损失函数 (置信度序列长度与标签序列长度相同), 对整个序列的损失函数求平均值, 得到模型的预测损失。基于预测损失可以通过反向传播的方式对音频处理模型进行训练。

步骤 503 中, 响应于播放触发操作, 在播放界面显示视频文件的推荐参数曲线。

示例的, 播放界面的推荐参数曲线与播放界面的时间轴的进度条绑定, 视频在播放界面播放时, 在进度条的上方显示精彩程度曲线, 精彩程度越高, 则曲线对应的数值越高, 用户可以根据精彩程度曲线拉取进度条, 定位到精彩数据片段进行观看。

5 本申请实施例提供的音频数据的处理方法的有益效果在于:

1、本申请实施例使用音频的信息来进行自动化的精彩数据片段识别, 自动化的定位能够快速工业化的定位出精彩数据片段, 在一些落地应用中, 尤其是像播放端的热度曲线(精彩程度曲线), 可以快速批量化的生产, 提升生产效率并降低生产成本。

2、本申请实施例使用全音频信息来做精彩数据片段定位的特征输入, 能够弥补画面非高燃但背景音乐高燃的数据片段定位不到的问题(比如情景剧), 尤其是使用画面来定位精彩数据片段只能定位出整个画面中最高燃的几个镜头, 无法完善整个精彩数据片段的完整性, 但是使用音频能够将整个数据片段定位出。并且普遍的画面处理模型参数量较大, 不能快速地预测出精彩数据片段, 音频网络参数较小, 更为快速便捷。

3、本申请实施例使用金字塔场景解析网络搭建信源分离系统, 然后再使用语音活动检测算法进行对语音段落定位。该方法能够将完全的语音进行检测, 不仅仅是语音信息, 能够让整个信源分离系统得知更完整的语音数据片段定位信息。

4、本申请实施例使用语音的时间段落信息来确定整个音轨中每个节点权重信息。本申请实施例能够直接定位出语音数据片段并对语音数据片段分配对应的权重值, 能够加强识别到语音数据片段的语义信息, 极大的提升了语音语义信息在精彩数据片段定位中的占比。

5、本申请实施例使用多域多层的方法来提取语义特征, 能够通过时域和频域在不同的网络层中互相补充信息, 在时域特征中添加了频域信息, 同样地, 在频域特征中添加了时域信息。使得高层网络感知到底层网络特征, 提升整个模型的感受域以及不同特征间的互补, 从而提升整个音频处理模型的定位性能。

下面继续说明本申请实施例提供的音频数据的处理装置 455 的实施为软件模块的示例性结构, 在一些实施例中, 如图 2 所示, 存储在存储器 450 的音频数据的处理装置 455 中的软件模块可以包括: 信源分离模块 4551, 配置为对音频数据进行信源分离, 得到至少一种信源类型分别对应的音轨数据; 权重配置模块 4552, 配置为基于每个音轨数据的播放时间轴中与信源类型相关的至少一个时间段落, 对音频数据中的每个数据片段基于所包含的所述时间段落长度分配对应的权重值, 并将每个权重值组合形成音频数据的权重值序列; 特征提取模块 4553, 配置为基于从每个数据片段提取的音频特征, 组合形成音频数据的音频特征序列, 并对音频特征序列进行编码, 得到音频数据的注意力参数序列; 参数预测模块 4554, 配置为对注意力参数序列与权重值序列进行融合处理, 得到每个数据片段的融合参数, 并基于每个融合参数确定每个数据片段的推荐参数; 参数预测模块 4554, 还配置为基于每个数据片段的推荐参数, 确定音频数据中的推荐片段。

在一些实施例中, 信源分离模块 4551, 配置为对音频数据进行特征提取处理, 得到音频数据的全局特征; 以全局特征为掩膜, 对音频数据进行信源分离, 得到音频数据中每种信源类型分别对应的音轨数据, 其中, 掩膜的边界用于表征不同信源类型对应的音频数据之间的边界。

在一些实施例中, 信源分离模块 4551, 配置为对音频数据进行特征提取处理, 得到音频数据的原始特征; 对原始特征进行多个层次的池化处理, 得到音频数据的多个局部特征; 对多个局部特征进行合并处理, 得到音频数据的全局特征。

在一些实施例中, 权重配置模块 4552, 配置为通过以下方式确定与信源类型相关的至少一个时间段落: 当音轨数据对应的信源类型为语音时, 将音轨数据中短时能量大于

能量阈值且过零率小于过零率阈值的时间段落,作为与语音相关的时间段落;当音轨数据对应的信源类型为背景音时,将音轨数据中满足筛选条件的时间段落作为与背景音相关的时间段落,其中,筛选条件包括以下任意一项:时间段落对应的响度大于响度下限值;时间段落的长度大于长度下限值。

5 在一些实施例中,权重配置模块 4552,配置为当通过信源分离得到语音和背景音两种信源类型的音轨数据时,针对每个数据片段进行以下处理:当数据片段属于语音相关的时间段落时,基于数据片段对应的语音的参数确定数据片段对应的权重值,其中,权重值与参数正相关,参数包括以下至少之一:语速、语调、响度;当数据片段属于背景音相关的时间段落时,将预设数值作为数据片段对应的权重值,其中,预设数值小于任意一个语音相关的数据片段的权重值;当数据片段不属于任意信源类型相关的时间段落时,将零作为数据片段对应的权重值。

10 在一些实施例中,权重配置模块 4552,配置为当通过信源分离仅得到背景音一种信源类型的音轨数据时,针对每个数据片段进行以下处理:当数据片段包含的时间段落属于背景音相关的时间段落时,基于数据片段对应的背景音的参数确定数据片段对应的权重值,其中,权重值与参数正相关,参数包括以下至少之一:响度、音调;当数据片段包含的时间段落不属于任意信源类型相关的时间段落时,将零作为数据片段对应的权重值。

15 在一些实施例中,特征提取模块 4553,配置为针对音频数据中每个数据片段进行以下处理:提取数据片段的时域信号特征与频域信号特征;基于每个音轨数据的播放时间轴中与信源类型相关的至少一个时间段落,确定时域信号特征对应的一维时域权重值,以及确定频域信号特征对应的二维频域权重值;对一维时域权重值与时域信号特征的乘积进行多个层次的卷积,得到时域音频特征;对二维频域权重值与频域信号特征的乘积进行多个层次的卷积,得到频域音频特征;对时域音频特征进行缩放,得到二维时域音频特征;对二维时域音频特征与频域音频特征进行融合处理,得到数据片段的音频特征。

20 在一些实施例中,特征提取模块 4553,配置为对二维时域音频特征与频域音频特征进行叠加处理,对叠加处理得到的叠加特征进行二维卷积,得到二维卷积结果,获取二维卷积结果的最大叠加特征与平均叠加特征;对最大叠加特征与平均叠加特征之间的加和进行线性激活,得到数据片段的音频特征。

25 在一些实施例中,参数预测模块 4554,配置为针对音频特征序列中每个音频特征执行以下处理:基于注意力机制对音频特征与其他数据片段的每个音频特征分别进行融合处理,得到音频特征对应的每个加权相关度;将每个加权相关度相加,得到音频特征对应的注意力参数,其中,其他数据片段是音频数据中除所述数据片段之外的数据片段;基于每个音频特征对应的数据片段的顺序,将每个注意力参数组合形成音频数据的注意力参数序列。

30 在一些实施例中,参数预测模块 4554,配置为在基于注意力机制对音频特征与其他数据片段的每个音频特征分别进行融合处理,得到音频特征对应的每个加权相关度之前,对音频数据的每个数据片段的音频特征进行全连接,得到每个音频特征的嵌入向量;针对音频特征与其他数据片段的每个音频特征执行以下处理:对音频特征的嵌入向量与其他数据片段的嵌入向量相乘,得到音频特征与其他数据片段的音频特征之间的相关度;将音频特征与相关度相乘,得到音频特征对应的加权相关度。

35 40 在一些实施例中,参数预测模块 4554,配置为针对每个数据片段执行以下处理:从注意力参数序列获取数据片段对应的注意力参数,将数据片段的权重值与数据片段的注意力参数相乘,得到数据片段的融合参数;对融合参数进行归一化处理,得到数据片段的推荐参数。

在一些实施例中，参数预测模块 4554，配置为通过以下任意一种方式确定音频数据的推荐片段：基于每个数据片段的推荐参数对每个数据片段进行降序排序，将降序排序的头部的至少一个数据片段作为音频数据的推荐片段；将推荐参数大于推荐参数阈值的数据片段作为推荐片段。

5 在一些实施例中，参数预测模块 4554，配置为基于每个数据片段的推荐参数，生成音频数据的推荐参数曲线；响应于播放触发操作，在播放界面显示音频数据的推荐参数曲线，其中，推荐参数曲线的横坐标为音频数据的播放时间，推荐参数曲线的纵坐标为推荐参数。

10 在一些实施例中，参数预测模块 4554，配置为在播放界面显示推荐片段的标签，其中，标签用于表征推荐片段的时间段落；响应于针对任意一个标签的选择操作，跳转到选中的标签对应的推荐片段的起点开始播放。

15 在一些实施例中，信源分离通过调用音频处理模型的金字塔场景解析网络实现，从每个数据片段提取音频特征通过调用音频处理模型的音频语义信息提取模块实现，编码与融合处理通过调用音频处理模型的注意力模块实现；其中，通过以下方式训练音频处理模型：基于音频数据的每个实际推荐片段的标签值，组合形成音频数据的实际推荐参数序列；基于音频数据的每个数据片段的推荐参数，组合形成音频数据的预测推荐参数序列；基于实际推荐参数序列与预测推荐参数序列获取音频处理模型的交叉熵损失；将交叉熵损失除以音频数据的数据片段数量，得到平均预测损失，基于平均预测损失对音频处理模型进行反向传播处理，得到更新后的音频处理模型。

20 本申请实施例提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令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指令，处理器执行该计算机指令，使得该电子设备执行本申请实施例上述的音频数据的处理方法。

25 本申请实施例提供一种存储有可执行指令的计算机可读存储介质，其中存储有可执行指令，当可执行指令被处理器执行时，将引起处理器执行本申请实施例提供的音频数据的处理方法，例如，如图 3A 示出的音频数据的处理方法。

30 在一些实施例中，计算机可读存储介质可以是 FRAM、ROM、PROM、EPROM、EEPROM、闪存、磁表面存储器、光盘、或 CD-ROM 等存储器；也可以是包括上述存储器之一或任意组合的各种设备。

35 在一些实施例中，可执行指令可以采用程序、软件、软件模块、脚本或代码的形式，按任意形式的编程语言（包括编译或解释语言，或者声明性或过程性语言）来编写，并且其可按任意形式部署，包括被部署为独立的程序或者被部署为模块、组件、子例程或者适合在计算环境中使用的其它单元。

40 作为示例，可执行指令可被部署为在一个电子设备上执行，或者在位于一个地点的多个电子设备上执行，又或者，在分布在多个地点且通过通信网络互连的多个电子设备上执行。

综上所述，通过本申请实施例从音频侧的角度来对整个视频进行多个域内以及多层信息的分析，能够快速定位出整个音频中的推荐片段（例如：精彩数据片段、热血数据片段、悲伤数据片段或者搞笑数据片段等），从而基于音频的推荐片段能够判断出视频中的推荐片段的时间段落在时间轴中的位置。从而在不依赖音频数据的播放记录数据的情况下，就可以准确识别推荐片段，为用户提供准确的参考信息，提升了用户体验。能够为播放器提供视频推荐参数曲线，以供观众将播放进度条由当前播放位置跳转到推荐片段的位置，提升观众对播放器的使用体验。

45 以上所述，仅为本申请的实施例而已，并非用于限定本申请的保护范围。凡在本申

请的精神和范围之内所作的任何修改、等同替换和改进等，均包含在本申请的保护范围之内。

权利要求书

1、一种音频数据的处理方法，由电子设备执行，所述方法包括：

从音频数据提取得到至少一种信源类型分别对应的音轨数据，其中，所述音频数据包含多个数据片段；

5 确定每个所述音轨数据的播放时间轴中与所述信源类型相关的至少一个时间段落，并确定所述音频数据中每个所述数据片段中分别包含的时间段落；

对所述音频数据中的每个数据片段基于所包含的所述时间段落长度分配对应的权重值，并将每个所述权重值组合形成所述音频数据的权重值序列；

10 从所述每个数据片段提取音频特征，将所述每个数据片段的音频特征组合形成所述音频数据的音频特征序列，并对所述音频特征序列进行编码，得到所述音频数据的注意力参数序列；

将所述注意力参数序列与所述权重值序列融合得到每个所述数据片段的融合参数，并基于每个所述融合参数确定每个所述数据片段的推荐参数；

基于每个所述数据片段的推荐参数，确定所述音频数据中的推荐片段。

15 2、如权利要求 1 所述的方法，其中，所述从音频数据提取得到至少一种信源类型分别对应的音轨数据，包括：

从所述音频数据提取得到所述音频数据的全局特征；

20 以所述全局特征为掩膜，对所述音频数据进行信源分离，得到所述音频数据中每种所述信源类型分别对应的音轨数据，其中，所述掩膜的边界用于表征不同信源类型对应的音频数据之间的边界。

3、如权利要求 2 所述的方法，其中，所述从所述音频数据提取得到所述音频数据的全局特征，包括：

对所述音频数据进行特征提取处理，得到所述音频数据的原始特征；

对所述原始特征进行多个层次的池化处理，得到所述音频数据的多个局部特征；

25 将所述多个局部特征合并得到所述音频数据的全局特征。

4、如权利要求 1 至 3 任一项所述的方法，其中，

所述确定每个所述音轨数据的播放时间轴中与所述信源类型相关的至少一个时间段落，并确定所述音频数据中每个所述数据片段中分别包含的时间段落，包括：

通过以下方式确定与所述信源类型相关的至少一个时间段落：

30 当所述音轨数据对应的信源类型为语音时，将所述音轨数据中短时能量大于能量阈值且过零率小于过零率阈值的时间段落，作为与所述语音相关的时间段落；

当所述音轨数据对应的信源类型为背景音时，将所述音轨数据中满足筛选条件的时间段落作为与所述背景音相关的时间段落，其中，所述筛选条件包括以下任意一项：所述时间段落对应的响度大于响度下限值；所述时间段落的长度大于长度下限值。

35 5、如权利要求 1 至 4 任一项所述的方法，其中，

当从所述音频数据提取得到语音和背景音两种信源类型的音轨数据时，所述对所述音频数据中的每个数据片段基于所包含的所述时间段落长度分配对应的权重值，包括：

针对每个所述数据片段进行以下处理：

40 当所述数据片段属于所述语音相关的所述时间段落时，基于所述数据片段对应的语音的参数确定所述数据片段对应的权重值，其中，所述权重值与所述参数正相关，所述参数包括以下至少之一：语速、语调、响度；

当所述数据片段属于所述背景音相关的所述时间段落时，将预设数值作为所述数据片段对应的权重值，其中，所述预设数值小于任意一个所述语音相关的数据片段的权重值；

当所述数据片段不属于任意所述信源类型相关的时间段落时，将零作为所述数据片段对应的权重值。

6、如权利要求 1 至 4 任一项所述的方法，其中，当从所述音频数据提取得到背景音一种信源类型的音轨数据时，所述对所述音频数据中的每个数据片段基于所包含的所述时间段落长度分配对应的权重值，包括：

针对每个所述数据片段进行以下处理：

当所述数据片段包含的时间段落属于所述背景音相关的所述时间段落时，基于所述数据片段对应的背景音的参数确定所述数据片段对应的权重值，其中，所述权重值与所述参数正相关，所述参数包括以下至少之一：响度、音调；

当所述数据片段包含的时间段落不属于任意所述信源类型相关的时间段落时，将零作为所述数据片段对应的权重值。

7、如权利要求 1 至 6 任一项所述的方法，其中，在所述将从所述每个数据片段提取的音频特征组合形成所述音频数据的音频特征序列之前，所述方法还包括：

针对所述音频数据中每个所述数据片段进行以下处理：

提取所述数据片段的时域信号特征与频域信号特征；

基于每个所述音轨数据的播放时间轴中与所述信源类型相关的至少一个时间段落，确定所述时域信号特征对应的一维时域权重值，以及确定所述频域信号特征对应的二维频域权重值；

对所述一维时域权重值与所述时域信号特征的乘积进行多个层次的卷积，得到时域音频特征；

对所述二维频域权重值与所述频域信号特征的乘积进行多个层次的卷积，得到频域音频特征；

对所述时域音频特征进行缩放，得到二维时域音频特征；

将所述二维时域音频特征与所述频域音频特征融合得到所述数据片段的音频特征。

8、如权利要求 7 所述的方法，其中，所述将所述二维时域音频特征与所述频域音频特征融合得到所述数据片段的音频特征，包括：

确定所述二维时域音频特征与所述频域音频特征的叠加特征，对所述的叠加特征进行二维卷积，得到二维卷积结果，获取所述二维卷积结果的最大叠加特征与平均叠加特征；

对所述最大叠加特征与所述平均叠加特征之间的加和进行线性激活，得到所述数据片段的音频特征。

9、如权利要求 1 至 8 任一项所述的方法，其中，所述将所述音频特征序列进行编码，得到所述音频数据的注意力参数序列，包括：

针对所述音频特征序列中每个所述音频特征执行以下处理：基于注意力机制，将所述音频特征与其他数据片段的每个所述音频特征分别融合得到所述音频特征对应的每个加权相关度；将每个所述加权相关度相加得到所述音频特征对应的注意力参数，其中，所述其他数据片段是所述音频数据中除所述数据片段之外的数据片段；

基于每个所述音频特征对应的数据片段的顺序，将每个所述注意力参数组合形成所述音频数据的注意力参数序列。

10、如权利要求 9 所述的方法，其中，在所述基于注意力机制，将所述音频特征与其他数据片段的每个所述音频特征分别融合得到所述音频特征对应的每个加权相关度之前，所述方法还包括：

对所述音频数据的每个所述数据片段的音频特征进行全连接，得到每个所述音频特征的嵌入向量；

所述基于注意力机制，将所述音频特征与其他数据片段的每个所述音频特征分别融合得到所述音频特征对应的每个加权相关度，包括：

针对所述音频特征与其他数据片段的每个所述音频特征执行以下处理：

对所述音频特征的嵌入向量与其他数据片段的嵌入向量相乘，得到所述音频特征与其他数据片段的音频特征之间的相关度；

将所述音频特征与所述相关度相乘，得到所述音频特征对应的加权相关度。

11、如权利要求 1 至 10 任一项所述的方法，其中，

所述将所述注意力参数序列与所述权重值序列融合得到每个所述数据片段的融合参数，包括：

针对每个所述数据片段执行以下处理：

从所述注意力参数序列获取所述数据片段对应的注意力参数，将所述数据片段的权重值与所述数据片段的注意力参数相乘，得到所述数据片段的融合参数；

所述基于每个所述融合参数确定每个所述数据片段的推荐参数，包括：

对每个所述数据片段的融合参数进行归一化处理，得到每个所述数据片段的推荐参数。

12、如权利要求 1 至 11 任一项所述的方法，其中，所述基于每个所述数据片段的推荐参数，确定所述音频数据中的推荐片段，包括：

通过以下任意一种方式确定所述音频数据的推荐片段：

基于每个所述数据片段的推荐参数对每个所述数据片段进行降序排序，将降序排序结果中从头部开始的至少一个数据片段作为所述音频数据的推荐片段；

将推荐参数大于推荐参数阈值的数据片段作为推荐片段。

13、如权利要求 1 至 12 任一项所述的方法，其中，

在所述将所述注意力参数序列与所述权重值序列融合得到每个所述数据片段的融合参数，并基于每个所述融合参数确定每个所述数据片段的推荐参数之后，所述方法还包括：

基于所述每个所述数据片段的推荐参数，生成所述音频数据的推荐参数曲线；

响应于播放触发操作，在播放界面显示所述音频数据的推荐参数曲线，其中，所述推荐参数曲线的横坐标为所述音频数据的播放时间，所述推荐参数曲线的纵坐标为所述推荐参数。

14、如权利要求 13 所述的方法，其中，在所述基于每个所述数据片段的推荐参数，确定所述音频数据中的推荐片段之后，所述方法还包括：

在所述播放界面显示所述推荐片段的标签，其中，所述标签用于表征所述推荐片段的时间段落；

响应于针对任意一个所述标签的选择操作，跳转到选中的所述标签对应的推荐片段的起点开始播放。

15、如权利要求 1 至 14 任一项所述的方法，其中，所述音频数据的处理方法通过调用音频处理模型实现，所述方法还包括：

通过以下方式训练所述音频处理模型：

基于所述音频数据的每个实际推荐片段的标签值，组合形成所述音频数据的实际推荐参数序列；

基于所述音频数据的每个所述数据片段的推荐参数，组合形成所述音频数据的预测推荐参数序列；

基于所述实际推荐参数序列与所述预测推荐参数序列获取所述音频处理模型的交叉熵损失；

将所述交叉熵损失除以所述音频数据的数据片段数量，得到平均预测损失，基于所述平均预测损失对所述音频处理模型进行反向传播处理，得到更新后的所述音频处理模型。

16、一种音频数据的处理装置，所述装置包括：

5 信源分离模块，配置为从音频数据提取得到至少一种信源类型分别对应的音轨数据；
权重配置模块，配置为基于每个所述音轨数据的播放时间轴中与所述信源类型相关的至少一个时间段落，对所述音频数据中的每个所述数据片段分配对应的权重值，并将每个所述权重值组合形成所述音频数据的权重值序列；

10 特征提取模块，配置为从所述每个数据片段提取音频特征，将所述每个数据片段的与音频特征组合形成所述音频数据的音频特征序列，并对所述音频特征序列进行编码，得到所述音频数据的注意力参数序列；

参数预测模块，配置为将所述注意力参数序列与所述权重值序列融合得到每个所述数据片段的融合参数，并基于每个所述融合参数确定每个所述数据片段的推荐参数；

15 所述参数预测模块，还配置为基于每个所述数据片段的推荐参数，确定所述音频数据中的推荐片段。

17、一种电子设备，所述电子设备包括：

存储器，用于存储可执行指令；

处理器，用于执行所述存储器中存储的可执行指令时，实现权利要求 1 至 15 任一项所述的方法。

20 18、一种计算机可读存储介质，存储有可执行指令，所述可执行指令被处理器执行时实现权利要求 1 至 15 任一项所述的方法。

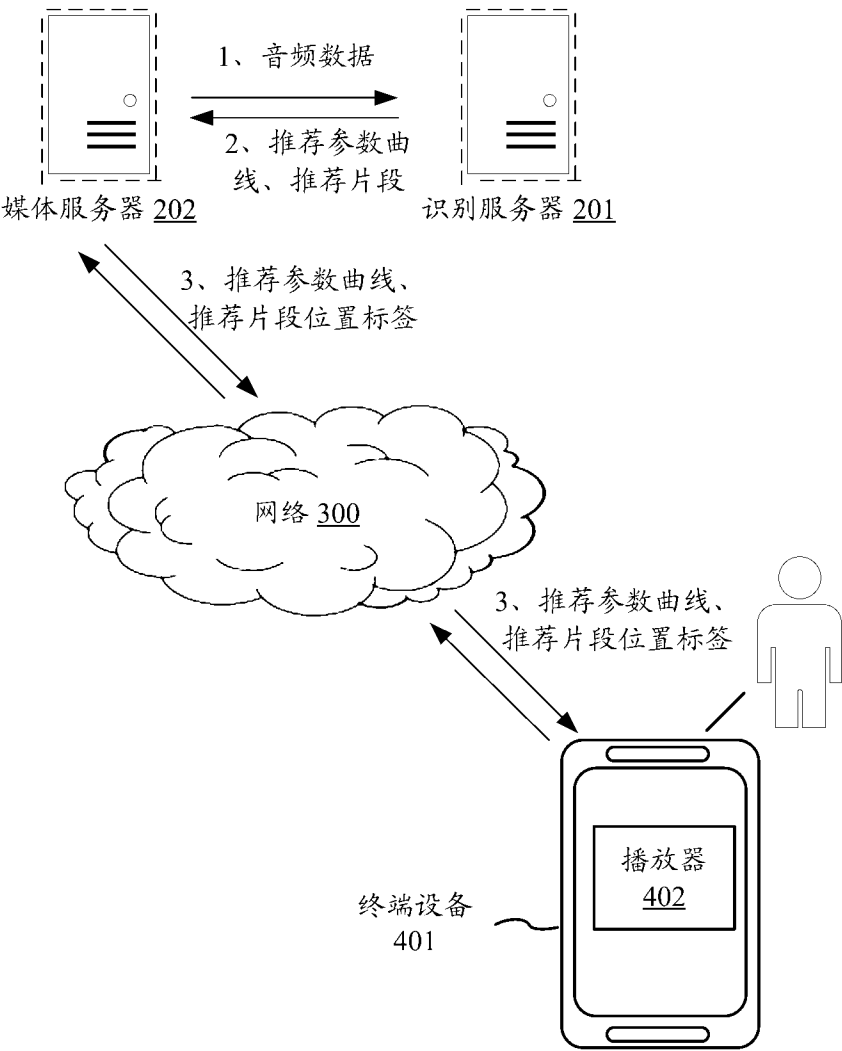


图 1

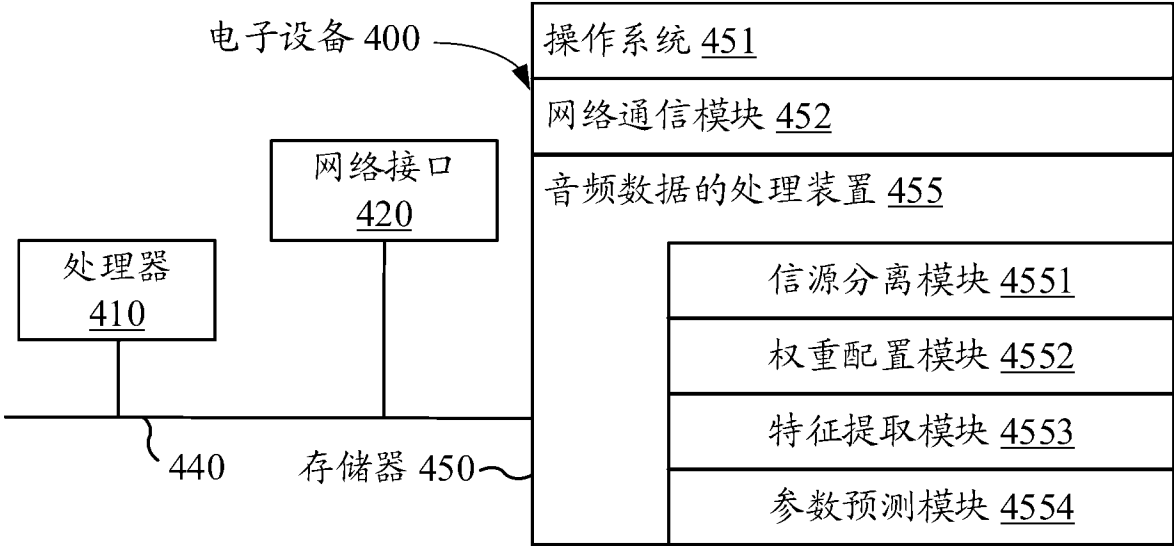


图 2

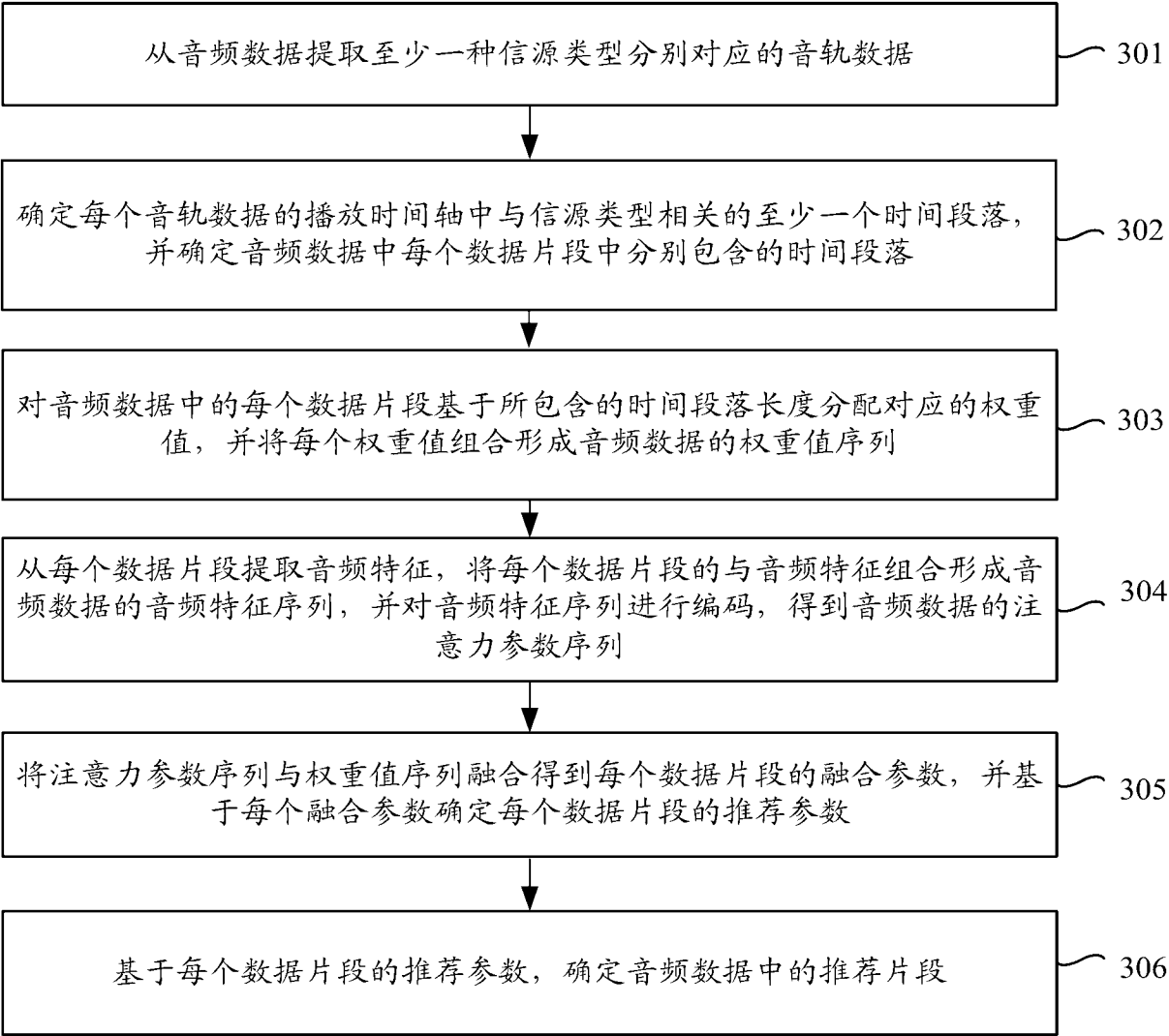


图 3A

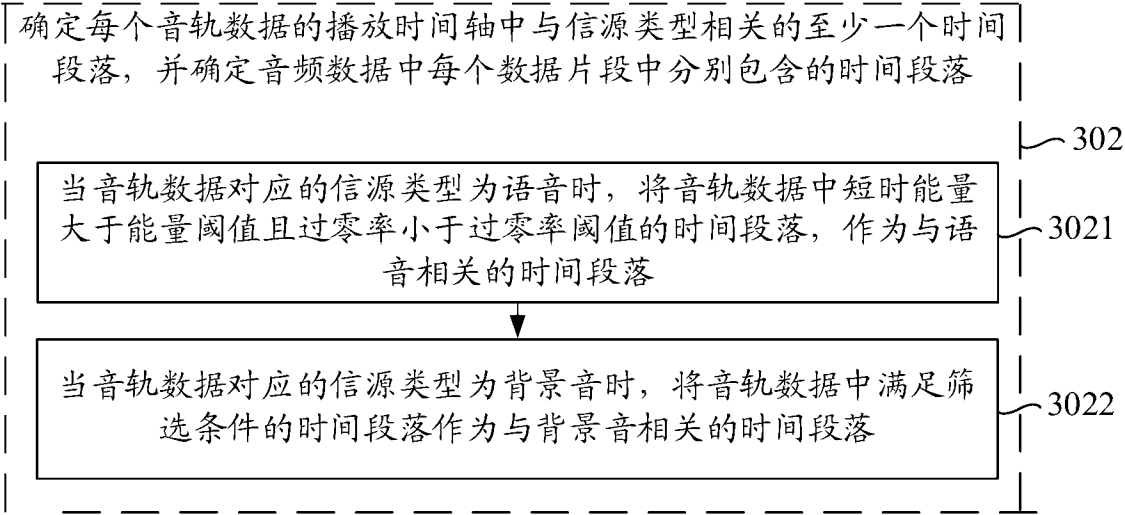


图 3B

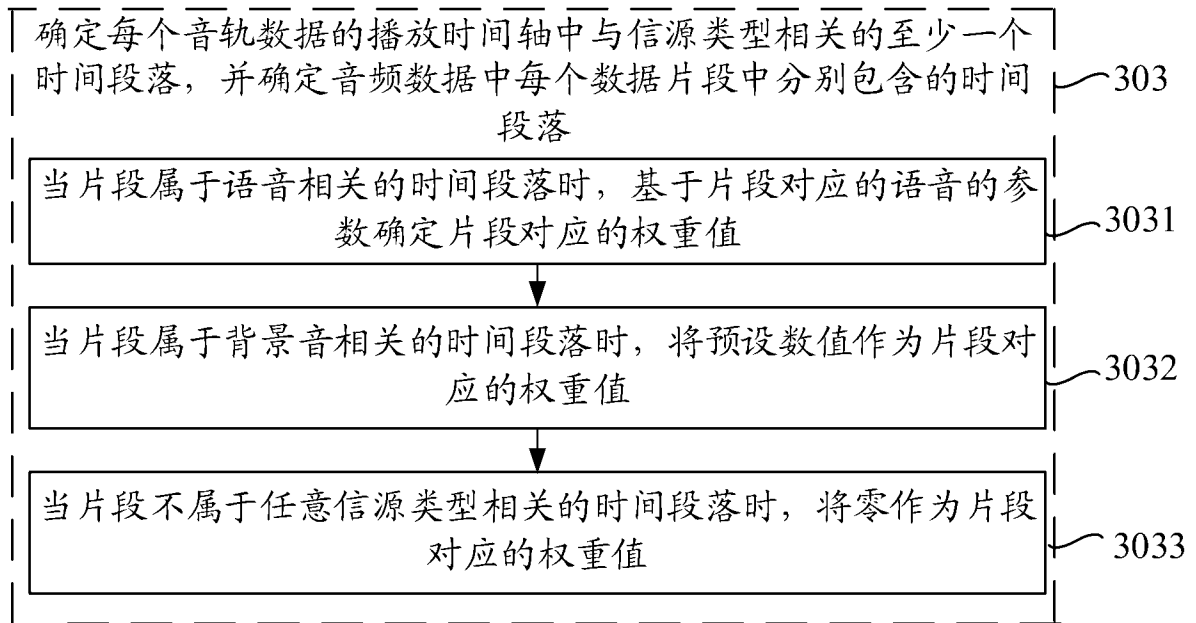


图 3C

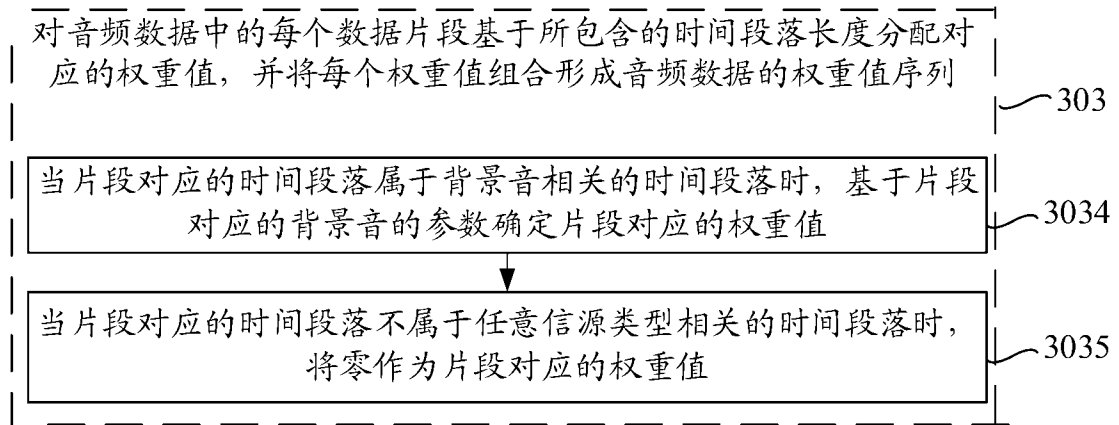


图 3D

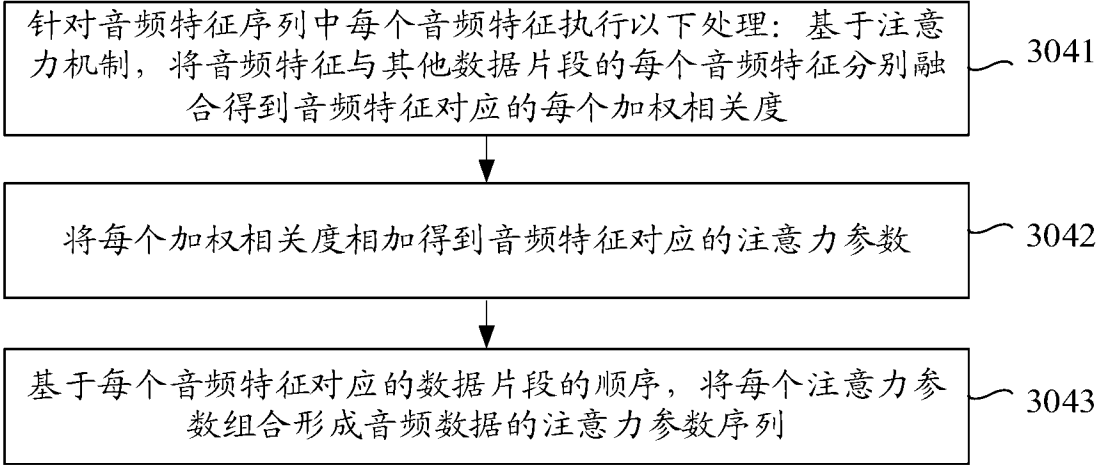


图 3E

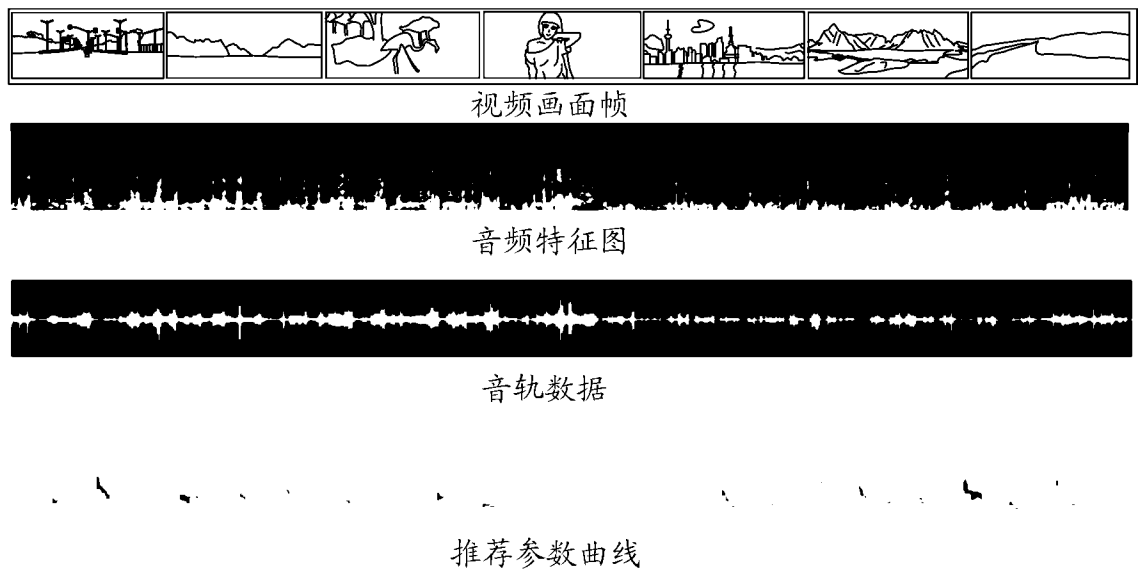


图 4A

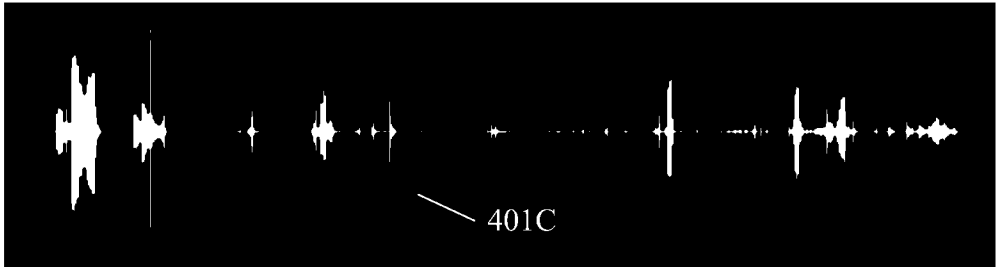
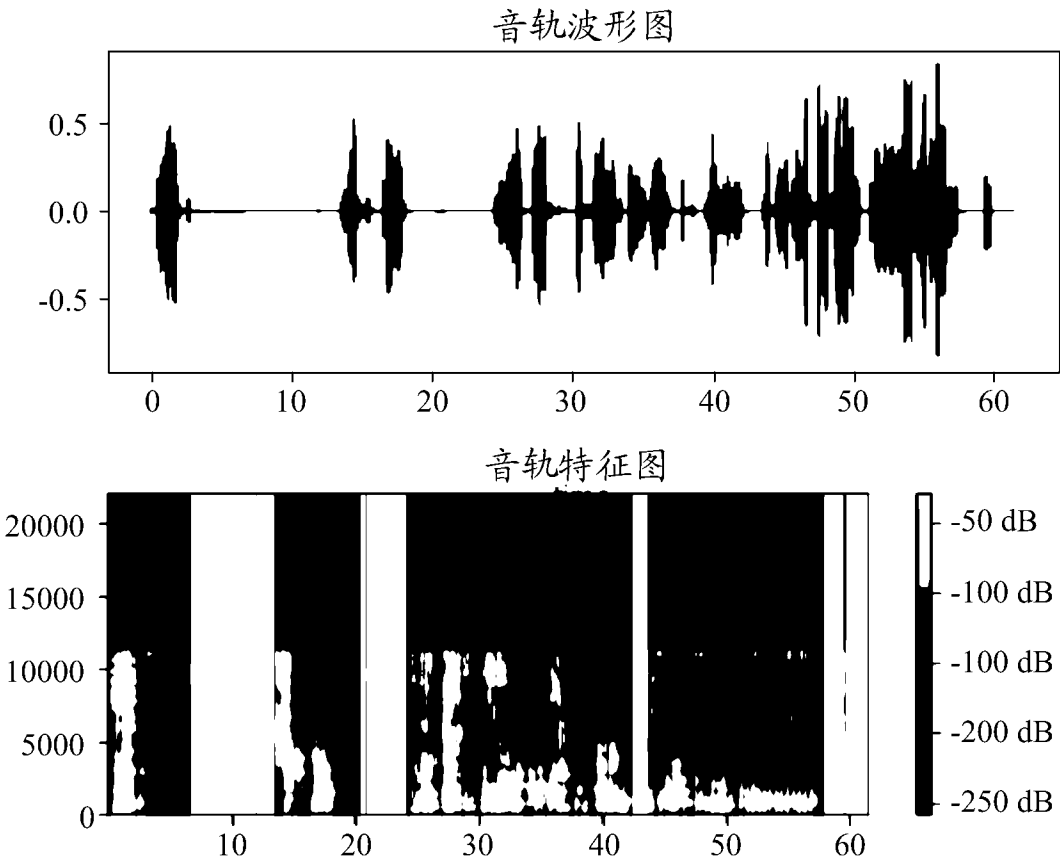


图 4C

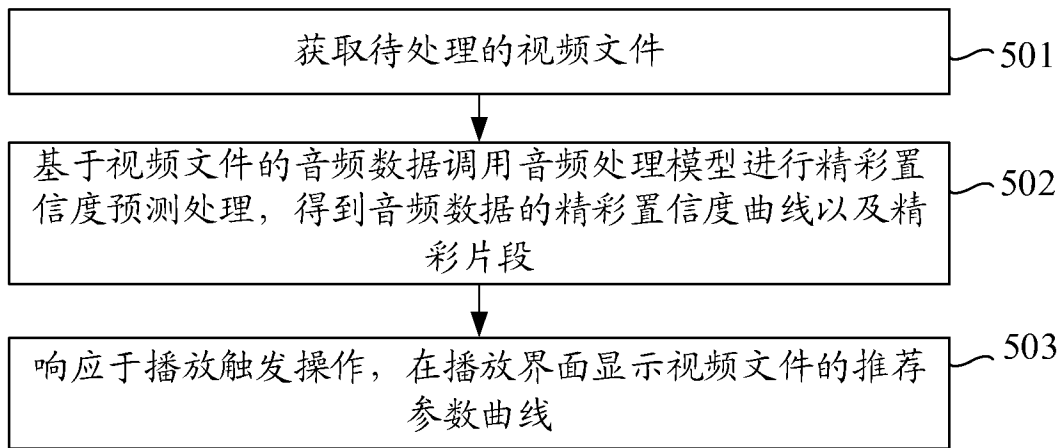


图 5

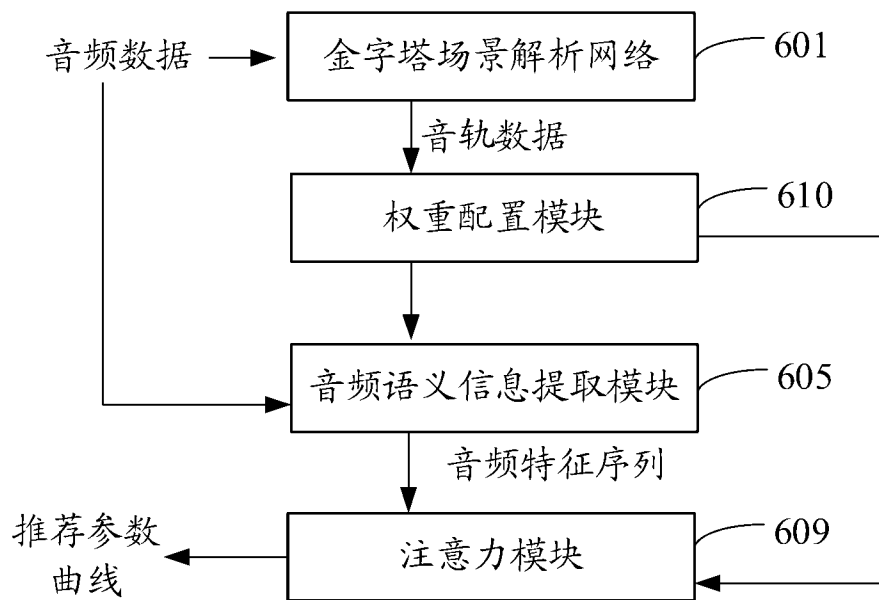


图 6A

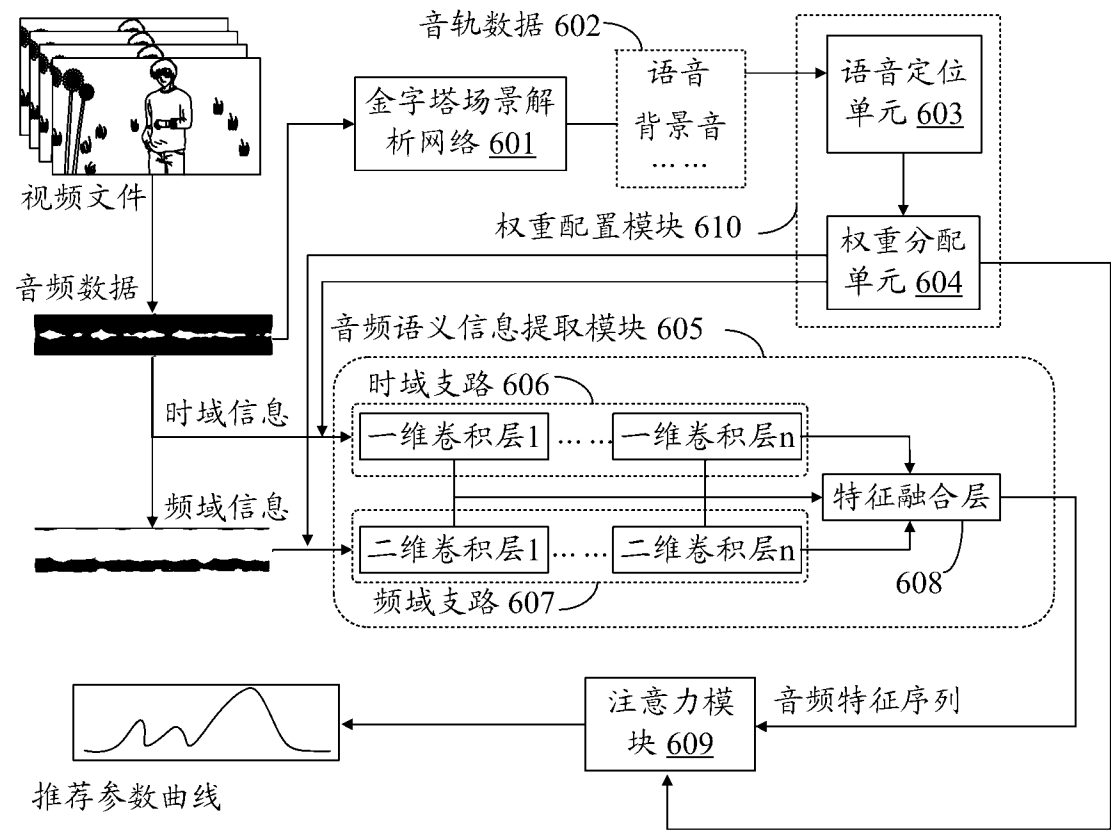


图 6B

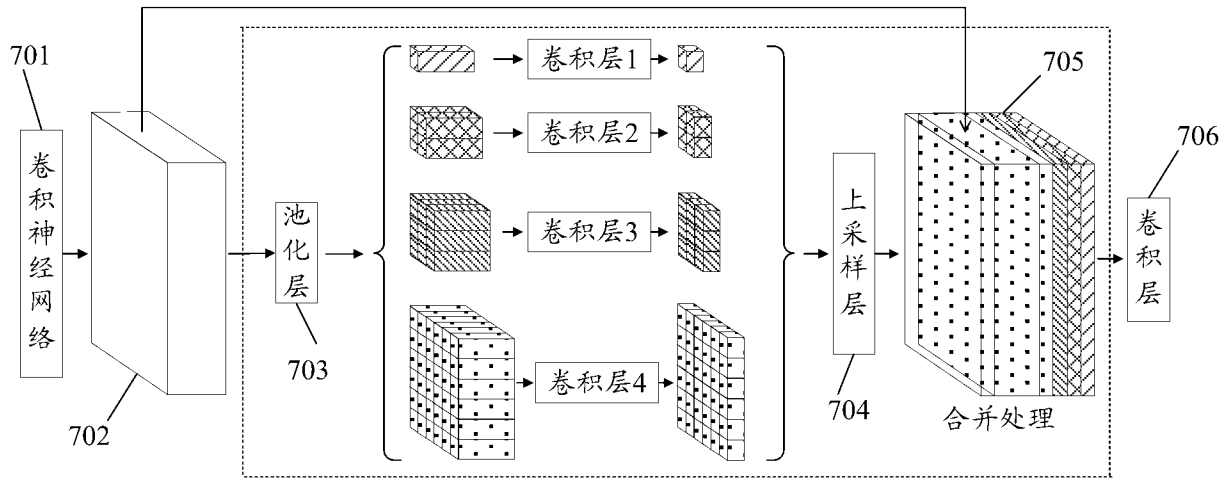


图 7

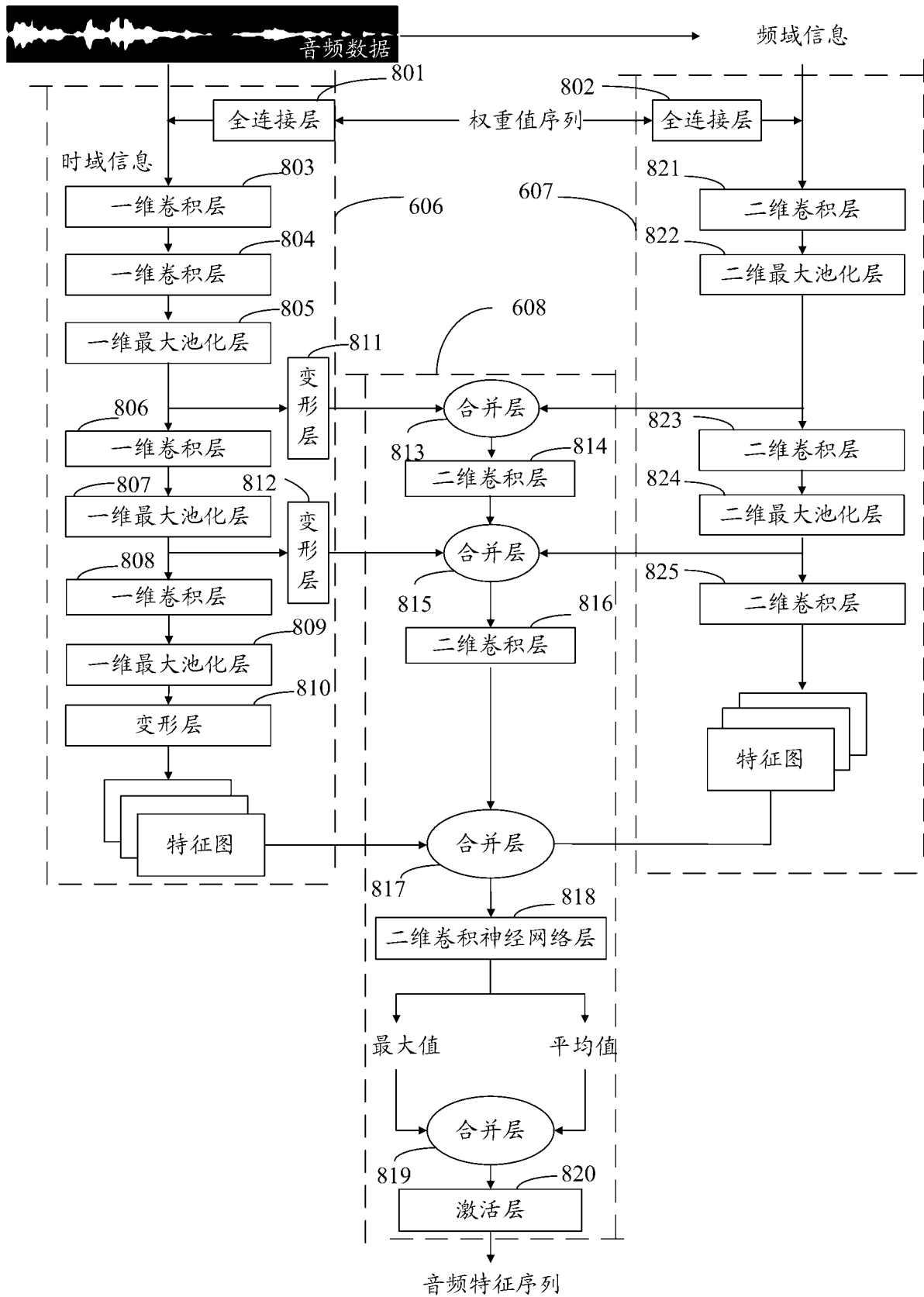
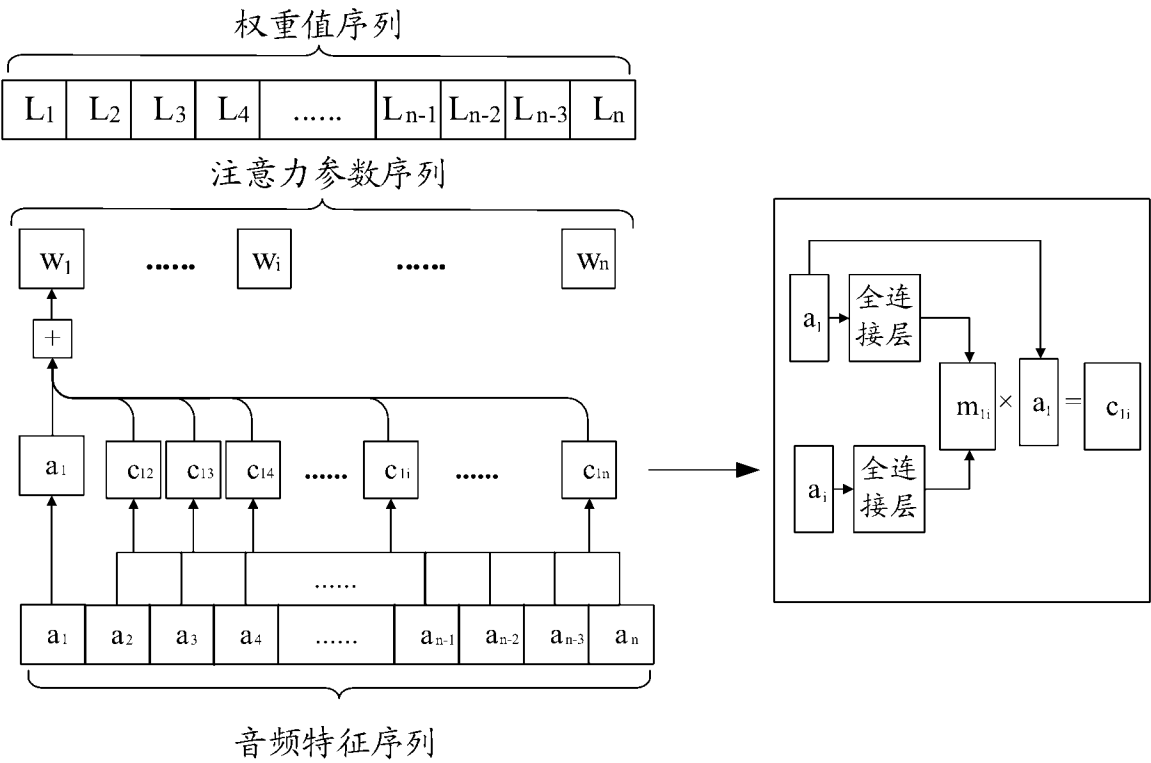


图 8



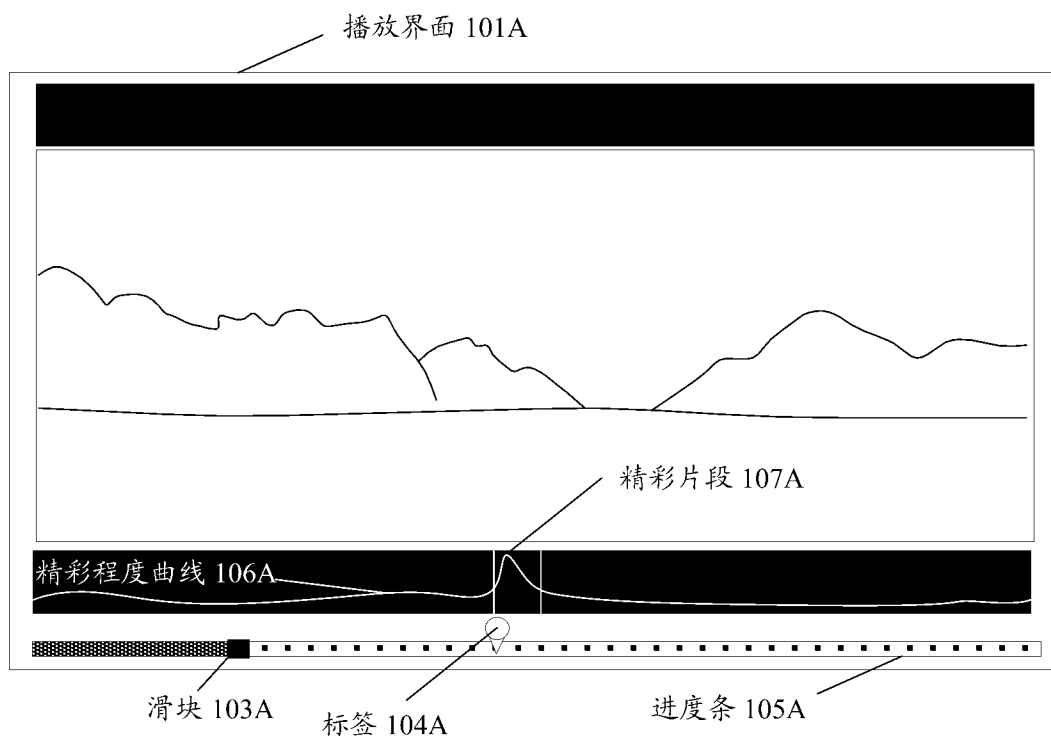


图 10A

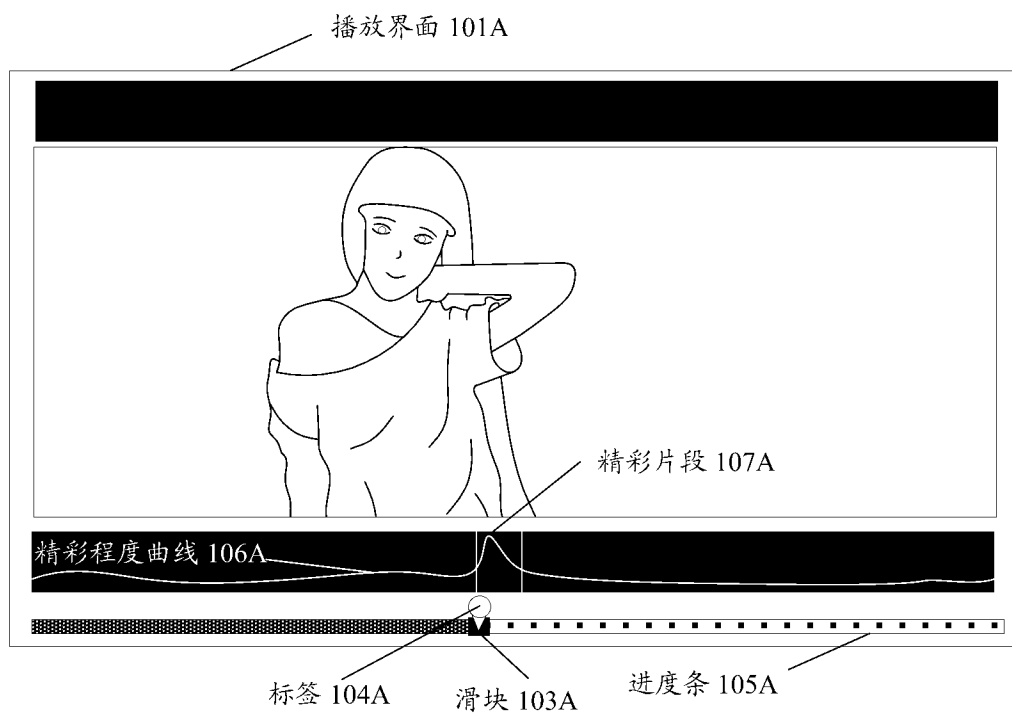


图 10B

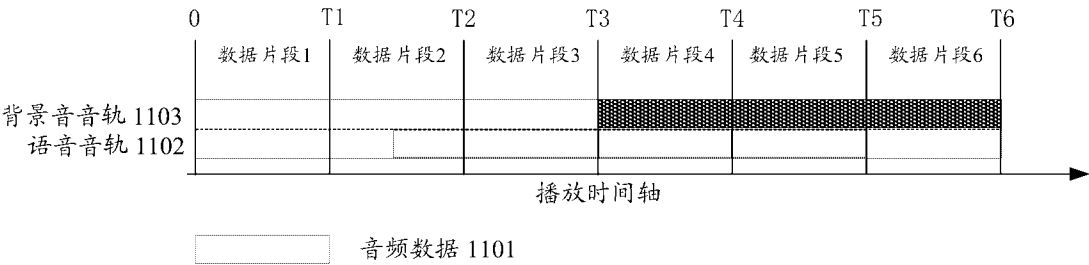


图 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/097205

A. CLASSIFICATION OF SUBJECT MATTER

G10L 15/02(2006.01)i; G10L 15/16(2006.01)i; G06N3/04(2023.01)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)

G10L; G06N

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)

CNTXT, ENTXTC, VEN, CNABS: 音频, 语音, 声音, 信源, 信号源, 类型, 音轨, 时间轴, 时间段, 权重, 权值, 加权, 特征, 序列, 编码, 注意力, 参数, 融合, 推荐, 标注, 提取, 精彩片段, 精彩内容, speech, voice, audio, sound, source, type, recommendation, time w axis, timeline, extract, feature?, segment+, weight, fuse, exciting, splendid, brilliant, wonderful, excerpt?, content

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 114822512 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 29 July 2022 (2022-07-29) claims 1-18	1-18
A	CN 112380377 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 19 February 2021 (2021-02-19) description, paragraphs [0026]-[0092], and figures 1-12	1-18
A	CN 101398826 A (SAMSUNG ELECTRONICS CO., LTD. et al.) 01 April 2009 (2009-04-01) entire document	1-17
A	CN 107154264 A (BEIJING DASHENG ONLINE TECHNOLOGY CO., LTD.) 12 September 2017 (2017-09-12) entire document	1-18
A	CN 111782863 A (TENCENT MUSIC ENTERTAINMENT TECHNOLOGY (SHENZHEN) CO., LTD.) 16 October 2020 (2020-10-16) entire document	1-18



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

22 August 2023

Date of mailing of the international search report

23 August 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/097205**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2021321172 A1 (SONY INTERACTIVE ENTERTAINMENT INC.) 14 October 2021 (2021-10-14) entire document	1-18
A	US 6973256 B1 (KONINKL PHILIPS ELECTRONICS N.V.) 06 December 2005 (2005-12-06) entire document	1-18

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/097205

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	114822512	A	29 July 2022	None			
CN	112380377	A	19 February 2021	None			
CN	101398826	A	01 April 2009	None			
CN	107154264	A	12 September 2017	None			
CN	111782863	A	16 October 2020	None			
US	2021321172	A1	14 October 2021	JP	2023521866	A	25 May 2023
				EP	4136857	A1	22 February 2023
				US	11381888	B2	05 July 2022
				WO	2021211368	A1	21 October 2021
US	6973256	B1	06 December 2005	WO	0237498	A2	10 May 2002
				WO	0237498	A3	22 August 2002
				KR	20020067581	A	22 August 2002
				KR	100833807	B1	30 May 2008
				EP	1332497	A2	06 August 2003
				JP	2004513547	A	30 April 2004

国际检索报告

国际申请号

PCT/CN2023/097205

A. 主题的分类 G10L 15/02(2006.01)i; G10L 15/16(2006.01)i; G06N3/04(2023.01)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) G10L; G06N 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)) CNTXT, ENTXTC, VEN, CNABS: 音频, 语音, 声音, 信源, 信号源, 类型, 音轨, 时间轴, 时间段, 权重, 权值, 加权, 特征, 序列, 编码, 注意力, 参数, 融合, 推荐, 标注, 提取, 精彩片段, 精彩内容, speech, voice, audio, sound, source, type, recommendation, time w axis, timeline, extract, feature?, segment+, weight, fuse, exciting, splendid, brilliant, wonderful, excerpt?, content		
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 114822512 A (腾讯科技(深圳)有限公司) 2022年7月29日 (2022 - 07 - 29) 权利要求1-18	1-18
A	CN 112380377 A (腾讯科技(深圳)有限公司) 2021年2月19日 (2021 - 02 - 19) 说明书第[0026]-[0092]段, 图1-12	1-18
A	CN 101398826 A (三星电子株式会社 等) 2009年4月1日 (2009 - 04 - 01) 全文	1-17
A	CN 107154264 A (北京大生在线科技有限公司) 2017年9月12日 (2017 - 09 - 12) 全文	1-18
A	CN 111782863 A (腾讯音乐娱乐科技(深圳)有限公司) 2020年10月16日 (2020 - 10 - 16) 全文	1-18
A	US 2021321172 A1 (SONY INTERACTIVE ENTERTAINMENT INC.) 2021年10月14日 (2021 - 10 - 14) 全文	1-18
☒ 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 2023年8月22日		国际检索报告邮寄日期 2023年8月23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		授权官员 张亚玲 电话号码 (+86) 62085722

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
A	US 6973256 B1 (KONINKL PHILIPS ELECTRONICS NV.) 2005年12月6日 (2005 - 12 - 06) 全文	1-18

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2023/097205

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	114822512	A	2022年7月29日	无			
CN	112380377	A	2021年2月19日	无			
CN	101398826	A	2009年4月1日	无			
CN	107154264	A	2017年9月12日	无			
CN	111782863	A	2020年10月16日	无			
US	2021321172	A1	2021年10月14日	JP	2023521866	A	2023年5月25日
				EP	4136857	A1	2023年2月22日
				US	11381888	B2	2022年7月5日
				WO	2021211368	A1	2021年10月21日
US	6973256	B1	2005年12月6日	WO	0237498	A2	2002年5月10日
				WO	0237498	A3	2002年8月22日
				KR	20020067581	A	2002年8月22日
				KR	100833807	B1	2008年5月30日
				EP	1332497	A2	2003年8月6日
				JP	2004513547	A	2004年4月30日