(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 11 月 21 日 (21.11.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/235271 A1

(51) 国际专利分类号:
G06T 13/40 (2011.01) *G06F 16/906* (2019.01)
G06T 13/20 (2011.01) *G06F 40/30* (2020.01)
G06F 16/907 (2019.01) *G06F 40/289* (2020.01)

(21) 国际申请号: PCT/CN2024/093505

(22) 国际申请日: 2024 年 5 月 15 日 (15.05.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202310547509.7 2023年5月15日 (15.05.2023) 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:卓嘉璇**(ZHUO, Jiaxuan)**; 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)。陆昱**(LU, Yu)**; 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)。付星辉**(FU, Xinghui)**; 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)。孙钟前**(SUN, Zhongqian)**; 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

WO 2024/235271 A1

(54) **Title:** MOVEMENT GENERATION METHOD AND APPARATUS FOR VIRTUAL CHARACTER, AND CONSTRUCTION METHOD AND APPARATUS FOR MOVEMENT LIBRARY OF VIRTUAL AVATAR

(54) 发明名称: 虚拟形象的动作生成方法、动作库的构建方法及装置

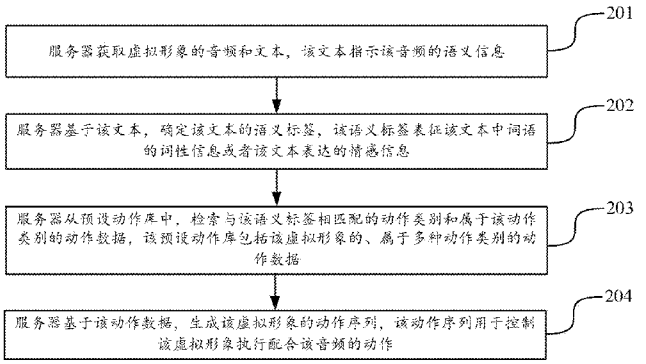

图 2

201 A server acquires audio and text of a virtual character, wherein the text indicates semantic information of the audio

202 The server determines a semantic label of the text on the basis of the text, wherein the semantic label represents at least one of part-of-speech information of words in the text and emotion information expressed in the text

203 The server retrieves, from a preset movement library, a movement category which matches the semantic label, and movement data which belongs to the movement category, wherein the preset movement library comprises movement data of the virtual character, which movement data belongs to a plurality of movement categories

204 The server generates a movement sequence for the virtual character on the basis of the movement data, wherein the movement sequence is used for controlling the virtual character to perform movements which match the audio

(57) **Abstract:** A movement generation method and apparatus for a virtual character, and a construction method and apparatus for a movement library of a virtual character, which methods and apparatuses belong to the technical field of computers. The movement generation method comprises: acquiring audio and text of a virtual character, wherein the text indicates semantic information of the audio (201); determining a semantic label of the text on the basis of the text, wherein the semantic label represents at least one of

[见续页]

淀区西直门北大街32号院1号楼4层503, Beijing 100088 (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, 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))。

part-of-speech information of words in the text and emotion information expressed in the text (202); retrieving, from a preset movement library, a movement category which matches the semantic label, and movement data which belongs to the movement category, wherein the preset movement library comprises movement data of the virtual character, which movement data belongs to a plurality of movement categories (203); and generating a movement sequence for the virtual character on the basis of the movement data, wherein the movement sequence is used for controlling the virtual character to perform movements which match the audio (204). The present application improves the efficiency of generating movement of a virtual character, and also improves the accuracy of movement generation.

(57) 摘要: 一种虚拟形象的动作生成方法、动作库的构建方法及装置，属于计算机技术领域。该方法包括：获取虚拟形象的音频和文本，文本指示音频的语义信息(201)；基于文本，确定文本的语义标签，语义标签表征文本中词语的词性信息或者文本表达的情感信息中的至少一项(202)；从预设动作库中，检索与语义标签相匹配的动作类别和属于动作类别的动作数据，预设动作库包括虚拟形象的、属于多种动作类别的动作数据(203)；基于动作数据，生成虚拟形象的动作序列，动作序列用于控制虚拟形象执行配合音频的动作(204)。本申请不但提升虚拟形象的动作生成效率，且提升动作生成准确率。

虚拟形象的动作生成方法、动作库的构建方法及装置

本申请要求于2023年05月15日提交、申请号为202310547509.7、发明名称为“虚拟形象的动作生成方法、动作库的构建方法及装置”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

本申请涉及计算机技术领域，特别涉及一种虚拟形象的动作生成方法、动作库的构建方法及装置。

背景技术

随着计算机技术的发展，虚拟形象在直播、影视、动漫、游戏、虚拟社交、人机交互等方面的应用越来越广泛。以直播场景为例，虚拟形象担任主播来进行播报或者对话，为了提升虚拟形象的渲染效果，涉及到虚拟形象的动作生成。

发明内容

本申请实施例提供了一种虚拟形象的动作生成方法、动作库的构建方法及装置，能够为虚拟形象快速、高效地合成准确率更高的动作序列，提升虚拟形象的动作生成效率。该技术方案如下：

一方面，提供了一种虚拟形象的动作生成方法，应用于计算机设备，所述方法包括：

获取虚拟形象的音频和文本，所述文本指示所述音频的语义信息；

基于所述文本，确定所述文本的语义标签，所述语义标签表征所述文本中词语的词性信息或者所述文本表达的情感信息中的至少一项；

从预设动作库中，检索与所述语义标签相匹配的动作类别和属于所述动作类别的动作数据，所述预设动作库包括所述虚拟形象的、属于多种动作类别的动作数据；

基于所述动作数据，生成所述虚拟形象的动作序列，所述动作序列用于控制所述虚拟形象执行配合所述音频的动作。

一方面，提供了一种虚拟形象的动作库的构建方法，应用于计算机设备，所述方法包括：

获取每个样本形象的样本动作序列、参考音频和参考文本，所述参考文本指示所述参考音频的语义信息，所述样本动作序列用于控制所述样本形象执行配合所述参考音频的动作；

基于所述参考文本中词语和所述参考音频中音素的关联关系，将所述样本动作序列划分为多个样本动作片段，每个样本动作片段与所述参考文本中的一个词语以及所述参考音频中的一个音素相关联；

基于所述样本动作片段的动作特征，对每个样本形象的每个样本动作片段进行聚类，得到多个动作集合，每个动作集合指示属于同一动作类别且属于不同样本形象的动作数据；

基于所述多个动作集合，构建动作库。

一方面，提供了一种虚拟形象的动作生成装置，所述装置包括：

获取模块，用于获取虚拟形象的音频和文本，所述文本指示所述音频的语义信息；

分析模块，用于基于所述文本，确定所述文本的语义标签，所述语义标签表征所述文本中词语的词性信息或者所述文本表达的情感信息中的至少一项；

检索模块，用于从预设动作库中，检索与所述语义标签相匹配的动作类别和属于所述动作类别的动作数据，所述预设动作库包括所述虚拟形象的、属于多种动作类别的动作数据；

生成模块，用于基于所述动作数据，生成所述虚拟形象的动作序列，所述动作序列用于控制所述虚拟形象执行配合所述音频的动作。

一方面，提供了一种虚拟形象的动作库的构建装置，所述装置包括：

样本获取模块，用于获取每个样本形象的样本动作序列、参考音频和参考文本，所述参考文本指示所述参考音频的语义信息，所述样本动作序列用于控制所述样本形象执行配合所述参考音频的动作；

片段划分模块，用于基于所述参考文本中词语和所述参考音频中音素的关联关系，将所述样本动作序列划分为多个样本动作片段，每个样本动作片段与所述参考文本中的一个词语以及所述参考音频中的一个音素相关联；

聚类模块，用于基于所述样本动作片段的动作特征，对每个样本形象的每个样本动作片段进行聚类，得到多个动作集合，每个动作集合指示属于同一动作类别且属于不同样本形象的动作数据；

构建模块，用于基于所述多个动作集合，构建动作库。

一方面，提供了一种计算机设备，该计算机设备包括一个或多个处理器和一个或多个存储器，该一个或多个存储器中存储有至少一条计算机程序，该至少一条计算机程序由该一个或多个处理器加载并执行以实现如上述任一种可能实现方式的虚拟形象的动作生成方法或虚拟形象的动作库的构建方法。

一方面，提供了一种计算机可读存储介质，该计算机可读存储介质中存储有至少一条计算机程序，该至少一条计算机程序由处理器加载并执行以实现如上述任一种可能实现方式的虚拟形象的动作生成方法或虚拟形象的动作库的构建方法。

一方面，提供一种计算机程序产品，所述计算机程序产品包括一条或多条计算机程序，所述一条或多条计算机程序存储在计算机可读存储介质中。计算机设备的一个或多个处理器能够从计算机可读存储介质中读取所述一条或多条计算机程序，所述一个或多个处理器执行所述一条或多条计算机程序，使得计算机设备能够执行上述任一种可能实施方式的虚拟形象的动作生成方法或虚拟形象的动作库的构建方法。

附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还能够根据这些附图获得其他的附图。

图 1 是本申请实施例提供的一种虚拟形象的动作生成方法的实施环境示意图；

图 2 是本申请实施例提供的一种虚拟形象的动作生成方法的流程图；

图 3 是本申请实施例提供的一种虚拟形象的动作生成方法的流程图；

图 4 是本申请实施例提供的一种虚拟形象的动作生成方法的原理图；

图 5 是本申请实施例提供的一种虚拟形象的动作库的构建方法的流程图；

图 6 是本申请实施例提供的一种动作库创建方法的原理图；

图 7 是本申请实施例提供的一种动作集合的数据清洗原理图；

图 8 是本申请实施例提供的一种新增动作片段的数据补充原理图；

图 9 是本申请实施例提供的一种虚拟形象的动作生成装置的结构示意图；

图 10 是本申请实施例提供的一种虚拟形象的动作库的构建装置的结构示意图；

图 11 是本申请实施例提供的一种计算机设备的结构示意图。

具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

本申请中术语“第一”“第二”等字样用于对作用和功能基本相同的相同项或相似项进行区分，应理解，“第一”、“第二”、“第 n”之间不具有逻辑或时序上的依赖关系，也不对数量和执行顺序进行限定。

本申请中术语“至少一个”是指一个或多个，“多个”的含义是指两个或两个以上，例如，多个动作片段是指两个或两个以上的动作片段。

本申请中术语“包括A或B中至少一项”涉及如下几种情况：仅包括A，仅包括B，以及包括A和B两者。

本申请中涉及到的用户相关的信息（包括但不限于用户的设备信息、个人信息、行为信息等）、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，当以本申请实施例的方法运用到具体产品或技术中时，均为经过用户许可、同意、授权或者经过各方充分授权的，且相关信息、数据以及信号的收集、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，本申请中涉及到的虚拟形象的动作数据都是在充分授权的情况下获取的。

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，它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。

人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习、自动驾驶、智慧交通等几大方向。

让计算机能听、能看、能说、能感觉，是未来人机交互的发展方向，其中语音成为未来最被看好的人机交互方式之一。语音技术（Speech Technology）的关键技术有自动语音识别（Automatic Speech Recognition，ASR）技术、语音合成技术、声纹识别技术等。

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。

自然语言处理（Nature Language Processing，NLP）是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

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，相信随着技术的发展，人工智能技术将在更多的领域得到应用，并发挥越来越重要的价值。

本申请实施例提供的方案涉及人工智能的语音技术、NLP和机器学习，具体涉及到利用以上各种技术或其组合在虚拟形象的动作生成方面的应用，将在以下的各个实施例中展开说明。

以下，对本申请实施例涉及的术语进行说明。

虚拟形象：一种在虚拟世界中可活动的对象，虚拟形象是虚拟世界中的一个虚拟的、拟人化的数字形象，如虚拟人物、动漫人物、虚拟角色等，虚拟形象可以是一个三维立体模型，该三维立体模型可以是基于三维人体骨骼技术构建的三维角色，可选地，虚拟形象也可以采

用2.5维或2维模型来实现，本申请实施例对此不加以限定。可以使用MMD（Miku Miku Dance，一种三维计算机图形软件）或者Unity引擎等来制作虚拟形象的3D模型，当然，也可以使用Live2D（一种二维计算机图形软件）来制作虚拟形象的2D模型，这里对虚拟形象的维度不进行具体限定。

元宇宙（Metaverse）：也称为后设宇宙、形上宇宙、超感空间、虚空间，是聚焦于社交链接的3D虚拟世界之网络，元宇宙涉及持久化和去中心化的在线三维虚拟环境。

数字人（Digital Human）：一种利用信息科学的方法对人体进行3D建模而生成的虚拟形象，达到对人体进行仿真、模拟的效果。再换一种表述，数字人是一种利用数字技术创造出来的、与人类形象接近的数字化人物形象。数字人广泛应用于视频创作、直播、行业播报、社交娱乐、语音提示等场景，例如，数字人可担任虚拟主播、虚拟化身等。其中，数字人也称为虚拟人、虚拟数字人等。

虚拟主播：指使用虚拟形象在视频网站上进行投稿活动的主播，例如虚拟YouTuber（Virtual YouTuber，VTuber）、虚拟UP主（Virtual Uploader，VUP）等。通常，虚拟主播以原创的虚拟人格设定、形象在视频网站、社交平台上进行活动，虚拟主播可以实现播报、表演、直播、对话等各种形式的人机交互。

中之人：指进行直播时候背后表演或操纵虚拟主播的人，比如，借助中之人安装在头部与肢体上的传感器，通过光学动作捕捉系统捕捉中之人的肢体动作和面部表情，将动作数据同步到虚拟主播上，这样能够借助实时运动捕捉的机制，实现虚拟主播与观看直播的观众之间的实时互动。

动作捕捉（Motion Capture，MoCap）：也称为运动捕捉。指代在运动物体或真人的关键部位上设置传感器，由动作捕捉系统捕捉传感器位置，再经过计算机处理后得到三维空间坐标的动作数据，当动作数据被计算机识别后，可以应用在动画制作、步态分析、生物力学、人机工程等领域。常见的动作捕捉设备包含动作捕捉服，多适用于3D虚拟形象的动作生成中，真人穿戴动作捕捉服来做出动作，从而将动作捕捉系统捕捉的人体3D骨架数据，迁移到虚拟形象的3D模型上，得到虚拟形象3D骨架数据，这一虚拟形象3D骨架数据将用于控制虚拟形象的3D模型执行与真人相同的动作。

光学动作捕捉：一种用于信息与系统科学相关工程与技术领域的仪器。

惯性动作捕捉：采用惯性传感器，可以对人体主要骨骼部位的运动进行实时测量，再根据反向运动学原理测算出人体关节的位置，并将数据施加到相应的（虚拟形象）骨骼上。

分词（Tokenization）：指将给定的一段文本，分解成以词语（Token）为单位的数据结构，每个词语中包含一个或多个字符。

词语（Token）：对一段给定的文本进行分词，将文本拆分成一个词语列表，词语列表中的每个元素就是一个分词得到的Token，每个Token中包含一个或多个字符。例如，将文本“我很开心”进行分词，将得到一个Token列表{“我”，“很”，“开心”}。

音素（Phone）：根据语音的自然属性划分出来的最小语音单位，依据音节里的发音动作来分析，一个动作构成一个音素。例如，一个词语中每个字符可能会按照发音动作被拆分成一个或多个音素。

音素对齐：给出一段音频和与该音频的语义对应的文本，将文本中每个字的音素拆分对齐到音频时间轴的每一个音频帧上。即，对文本中每个字，按照这个字的发音动作确定一个或多个音素，再从音频中找到发出每个音素的一个或多个音频帧，这样说出这个字需要发出的所有音素覆盖的所有音频帧构成一个音频片段，在音频时间轴上找到这个音频片段的时间戳区间，就反映了在音频中的哪个时间戳区间里说话人在说这个字。

插帧：是一种运动预估及运动补偿方式，能够在帧数不足的情况下扩展动作片段的动作帧数，使得动作变得连贯。例如，在动作片段原有的每两个动作帧中插入一个新的动作帧，利用新的动作帧来补充以上两个动作帧中动作变化的中间状态。

文本情感分析：给定一段文本，对该文本进行分析、处理、归纳和推理的过程，通常会输出与该文本匹配度最高的情感标签，因此又称为意见挖掘、倾向性分析。按照处理文本的粒度不同，情感分析大致可分为词语级、句子级、篇章级三个研究层次。文本情感分析的途径大致可以集合成四类：关键词识别、词汇关联、统计方法和概念级技术。

以下，对本申请实施例的技术构思进行说明。

随着三维（3-Dimension，3D）建模、虚拟现实（Virtual Reality，VR）、增强现实（Augmented Reality，AR）、元宇宙等技术的飞速发展，虚拟形象在直播、影视、动漫、游戏、虚拟社交、人机交互等方面的应用越来越广泛。

以直播场景为例，虚拟形象担任主播来进行播报或者对话，为了提升虚拟形象的渲染效果，涉及到虚拟形象的动作生成。同理，在视频创作场景下，如创作虚拟主播的投稿视频、创作数字人视频等，同样也会涉及到虚拟形象的动作生成。

通常，在生成虚拟形象的肢体动作时，采用动作捕捉方式：由真人（或称为演员）穿戴具有全身传感器的动作捕捉服，真人根据台本内容和台本音频进行动作表演，动作捕捉服捕捉真人表演的动作数据（即人体 3D 骨架数据），上报给动作捕捉服联机的计算机，计算机将人体 3D 骨架数据迁移到虚拟形象的 3D 模型上，得到虚拟形象 3D 骨架数据。此后，由于连续时刻下虚拟形象 3D 骨架数据将形成一段动作序列，再由专业的动画师在虚拟形象的动作序列上进行一些抖动或者矫正，对虚拟形象进行动作修复，最终得到一系列台本下所应具有的虚拟形象动作表现。以上基于动作捕捉的动作生成方式，整个过程需要人工干预，且每次捕捉的人体 3D 骨架数据都是根据特定台本而定制化的，不可重复利用，不具有通用性，即，一旦出现一段台本没有的音频或文本，那么将无法实现动作生成，需要演员以这段新的音频或文本作为新的台本来进行表演，因此动作生成效率低。

再者，在生成虚拟形象的肢体动作时，还可以通过大量公开的 2D 视频素材（如演讲视频、脱口秀视频等）来进行视频动作捕捉，得到 2D 视频数据，再转换成 3D 骨骼数据，以 3D 骨骼数据和其标注的音频及文本构建训练数据集，来训练一个动作生成模型，使得动作生成模型能够在音频驱动下生成虚拟形象的肢体动作。然而，由于数据来源较为单一，且人体动作复杂性很高，因此动作生成模型的效果不理想，最终合成的虚拟形象存在肢体动作趋于平淡、表演不准确等问题，因此动作生成准确率差。

有鉴于此，本申请实施例提出一种虚拟形象的动作库的构建方法，能够根据采集到的大量样本对象的样本动作序列及其参考文本、参考音频，根据参考文本和参考文本，来将样本动作序列分成样本动作片段，再将样本动作片段匹配到所属的动作类别，进而对每个动作类别的动作集合进行动作数据的清洗、过滤等，最终建立一个较为完善的虚拟形象的动作库，这一动作库能够覆盖较多的动作类别。接着，基于建立完毕的动作库，能够提供一种音频触发的肢体动作生成算法框架，在实时生成虚拟形象动作时，用户只需要给定一段音频和其释义的文本，就能够由机器快速实现音频和文本触发的肢体动作 3D 数据的生成，输出虚拟形象的动作序列，整个动作生成过程无需人工介入，机器能够快速、精确的生成配合音频和文本的动作序列，其动作生成效率高、动作生成准确率高。

如果不考虑文本模态的语义信息，仅根据与输入音频的相似性来查找动作片段以合成动作序列，那么最终肢体动作只会根据音频节奏简单变化，而无法反映出来真实语义层面的肢体动作，而且只能简单重复对白动作效果，无法表现出来语义准确度和丰富度，这显然动作生成效果不好，虚拟形象仿真度差。

但在以上技术方案中，由于考虑了音频、文本双模态的信息，来驱动虚拟形象的肢体动作生成，结合文本和音频，考虑两者之间的关联关系，在文本语义的指导下从预设动作库中匹配到丰富的语义动作，这样能够使得合成的动作序列中，虚拟形象的肢体动作表现更加准确、丰富和生动，可适用于各类虚拟形象需要执行动作的场景，如虚拟直播、数字人视频等场景，达到动作捕捉级别的准确率，但动作生成效率远优于动作捕捉方式。

以下，对本申请实施例的实施环境进行说明。

图 1 是本申请实施例提供的一种虚拟形象的动作生成方法的实施环境示意图。参见图 1，该实施环境包括终端 101 和服务器 102。终端 101 和服务器 102 之间通过无线网络或者有线网络进行直接或间接地连接，本申请在此不做限制。

终端 101 上安装有支持虚拟形象的应用，终端 101 能够通过该应用实现虚拟形象的肢体动作生成等功能，当然，该应用还能够具有其他功能，例如，网络社交功能、视频分享功能、视频投稿功能或者聊天功能等。其中，该应用为终端 101 操作系统中的原生应用，或者为第三方提供的应用。例如，该应用包括但不限于：直播应用、短视频应用、音视频应用、游戏应用、社交应用、3D 动画应用、或者其他应用，本公开实施例对此不做限制。

可选地，终端 101 是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等，但并不局限于此。

服务器 102 为终端 101 上支持虚拟形象的应用提供后台服务，服务器 102 创建并维护虚拟形象的动作库，并缓存有多种虚拟形象的 3D 骨骼模型。服务器 102 包括一台服务器、多台服务器、云计算平台或者虚拟化中心中的至少一种。可选地，服务器 102 承担主要动作生成计算工作，终端 101 承担次要动作生成计算工作；或者，服务器 102 承担次要动作生成计算工作，终端 101 承担主要动作生成计算工作；或者，服务器 102 和终端 101 之间采用分布式计算架构进行协同动作生成计算。

可选地，服务器 102 是独立的物理服务器，或者是多个物理服务器构成的服务器集群或者分布式系统，或者是提供云服务、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（Content Delivery Network，内容分发网络）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 101 可以泛指多个终端中的一个，本公开实施例仅以终端 101 来举例说明。本领域技术人员可以知晓，上述终端的数量可以更多或更少。

在一个示例性场景中，在实时肢体动作生成过程中，用户在终端 101 的应用中上传一段音频，触发动作生成指令，终端 101 响应于该动作生成指令，向服务器 102 发送动作生成请求，动作生成请求携带该音频。服务器 102 响应于该动作生成请求，对该音频进行自动语音识别（ASR），得到指示该音频语义的文本，接着，利用该音频和该文本，来执行本申请实施例涉及的虚拟形象的动作生成方法，从预设动作库中检索到合适的动作数据，进而合成配合该音频的动作序列。这样，终端 101 侧实现了音频驱动下的虚拟形象动作生成，但在服务器 102 侧则会利用音频和文本的双模态信息，来合成能够表征音频（或文本）语义层面的虚拟形象肢体动作。

在另一个示例性场景中，在实时肢体动作生成过程中，用户在终端 101 的应用中上传一段文本，触发动作生成指令，终端 101 响应于该动作生成指令，向服务器 102 发送动作生成请求，动作生成请求携带该文本。服务器 102 响应于该动作生成请求，找到虚拟形象的音源库，从该音源库中为该文本生成一段念出该文本的音频（即为文本配音），接着，利用该音频和该文本，来执行本申请实施例涉及的虚拟形象的动作生成方法，从预设动作库中检索到合适的动作数据，进而合成配合该文本的动作序列。这样，终端 101 侧实现了文本驱动下的虚拟形象动作生成，但在服务器 102 侧则会利用音频和文本的双模态信息，来合成能够表征音频（或文本）语义层面的虚拟形象肢体动作。

在又一个示例性场景中，在实时肢体动作生成过程中，用户在终端 101 的应用中上传一段音频和其对应的文本（即表征音频的语义信息的文本），触发动作生成指令，终端 101 响应于该动作生成指令，向服务器 102 发送动作生成请求，动作生成请求携带该音频和该文本。服务器 102 响应于该动作生成请求，利用该音频和该文本，来执行本申请实施例涉及的虚拟形象的动作生成方法，从预设动作库中检索到合适的动作数据，进而合成配合该音频和该文

本的动作序列。这样，终端 101 侧实现了音频和文本共同驱动下的虚拟形象动作生成，但在服务器 102 侧则会利用音频和文本的双模态信息，来合成能够表征音频（或文本）语义层面的虚拟形象肢体动作。

在以上各类场景中，不论终端 101 侧提供的是单模态信息还是双模态信息作为驱动信号，利用语音技术中的文字与音频之间的转换手段，在服务器 102 侧都会利用音频和文本的双模态信息来合成虚拟形象肢体动作，使得最终动作序列不但能够贴合音频节奏进行律动，而且能够表达语义层面的丰富语义信息，甚至体现出虚拟形象播报时的情感状态，因此不但动作生成效率高，而且动作生成准确率高，生成的肢体动作与音频节奏配合好、携带丰富语义信息，使得虚拟形象的仿真度也大幅提升，渲染效果得到极大优化。

本申请实施例提供的虚拟形象的动作生成方法，能够适用于任意需要生成虚拟形象肢体动作的场景下。例如，数字人直播场景下，中之人不需要配备动作捕捉服来进行表演，只需要给定直播互动时的文本或音频中至少一项，就能够控制数字人在音频和文本双模态信息的驱动下，做出直播中配合音频及其字幕（也可能没有字幕）的肢体动作，提升了数字人直播的真实性和趣味性。又例如，在创作数字人视频的场景下，用户只需要创作好视频的音频或文本，就能够控制生成与音频或文本配合的数字人肢体动作，进而将肢体动作（即视频画面）和音频（即视频配音）合成一段数字人视频，从而进行视频投稿、视频发布等，提升了数字人视频的生成效率，提升其创作便捷性、灵活性。又例如，还可以适用于数字人客服、动画制作、影视特效、数字人主持等各类需要生成虚拟形象肢体动作的场景，本申请实施例对应用场景不进行具体限定。

以下，对本申请实施例的虚拟形象的动作生成方法的流程进行说明。

图 2 是本申请实施例提供的一种虚拟形象的动作生成方法的流程图。参见图 2，该实施例由计算机设备执行，以计算机设备为服务器为例进行说明，服务器可以为上述实施环境的服务器 102，该实施例包括以下步骤。

201、服务器获取虚拟形象的音频和文本，该文本指示该音频的语义信息。

其中，虚拟形象是指一种在虚拟世界中可活动的对象，虚拟形象是虚拟世界中的一个虚拟的、拟人化的数字形象，例如，虚拟形象包括但不限于：游戏人物、虚拟主播、虚拟化身、影视人物、动漫人物、数字人、虚拟人等，本申请实施例对虚拟形象不进行具体限定。

本申请实施例中，在需要控制虚拟形象播报音频的情况下，还需要控制虚拟形象执行配合该音频的动作，因此服务器会生成虚拟形象的动作序列。

其中，音频中包含至少一个音频帧，文本则是指示该音频的语义信息的文本，文本中包含至少一个词语，每个词语包含至少一个字符。音频和文本具有关联关系，即文本是对音频进行 ASR 识别到的语义信息，或者音频是播报该文本发出的语音信号，语音信号可以是机器输出的合成信号，也可以是麦克风采集的人声信号，这里对语音信号的类型不进行具体限定。

在一些实施例中，服务器从本地数据库中查询到一对具有关联关系的音频和文本，或者，服务器从本地数据库中取出一段音频，对该音频进行 ASR 识别，得到指示该音频的语义信息的文本，又或者，服务器从本地数据库中取出一段文本，对该文本进行声音合成，得到为该文本配音的音频。

在另一些实施例中，服务器从云端数据库中下载一对具有关联关系的音频和文本，或者，服务器从云端数据库中下载一段音频，对该音频进行 ASR 识别，得到指示该音频的语义信息的文本，又或者，服务器从云端数据库中下载一段文本，对该文本进行声音合成，得到为该文本配音的音频。

在又一些实施例中，服务器接收终端上传的一对具有关联关系的音频和文本，例如，终端向服务器发送动作生成请求，服务器接收并解析该动作生成请求，得到音频和文本。或者，服务器接收终端上传的音频，对该音频进行 ASR 识别，得到指示该音频的语义信息的文本，例如，终端向服务器发送动作生成请求，服务器接收并解析该动作生成请求，得到音频，对

该音频进行 ASR 识别，得到指示该音频的语义信息的文本。或者，服务器接收终端上传的文本，对该文本进行声音合成，得到为该文本配音的音频，例如，终端向服务器发送动作生成请求，服务器接收并解析该动作生成请求，得到文本，对该文本进行声音合成，得到为该文本配音的音频。

在以上过程中，由于文本和音频可以互相转换，因此用户可以仅给定音频，也可以仅给定文本，还可以同时给定音频和文本，除了用户指定以外，也可以从本地数据库读取或者从云端数据库下载，本申请实施例对该音频和该文本的来源不进行具体限定。

服务器获取到音频和文本后，执行本申请实施例提供的方法生成动作序列，该动作序列与文本的语义信息匹配，则后续在控制该虚拟形象播报该音频的过程中，控制该虚拟对象执行该动作序列所指示的肢体动作，以使该虚拟对象执行的肢体动作的语义信息与所播报的音频匹配。

202、服务器基于该文本，确定该文本的语义标签，该语义标签表征该文本中词语的词性信息或者该文本表达的情感信息。

在一些实施例中，服务器对步骤 201 中获取到的文本进行分析，以获取到该文本的至少一个语义标签，其中，该语义标签可以包括词性标签或者情感标签中的至少一项，该词性标签表征该文本中词语的词性信息，词语的词性信息是指用于描述词语的词性的信息，如主语、动词、状态等，该情感标签表征该文本所表达的情感信息，情感信息是指用于描述该文本所表达的情感的信息，如高兴、失落、愤怒等，词性信息和情感信息都是对该文本进行描述，但是所描述的角度不同，本申请实施例对语义标签的内容不进行具体限定。该语义标签的数量可以是一个或多个，本申请实施例对语义标签的数量也不进行具体限定。

在一些实施例中，服务器基于该文本，确定该文本中包含的至少一个词语，对每个词语确定该词语所属的词性标签，将该文本中全部词语的词性标签作为该文本的语义标签。关于提取词性标签的方式将在下一实施例详细说明，此处不再赘述。

在另一些实施例中，服务器基于该文本，确定该文本的至少一个情感标签，将该至少一个情感标签作为该文本的语义标签。关于提取情感标签的方式将在下一实施例详细说明，此处不再赘述。

在又一些实施例中，服务器既基于该文本确定每个词语的词性标签，又基于该文本确定该文本的每个情感标签，接着，将每个词性标签和每个情感标签一起作为该文本的语义标签。

在一个示例中，针对文本“我第一次直播！”进行分词，得到词语列表{“我”，“第一次”，“直播！”}，其中，在词性表中查询到词语“我”所属的词性标签为“主语”，词语“第一次”所属的词性标签为“状态”，词语“直播！”所属的词性标签为“动词”，另外，基于该文本确定该文本的情感标签“高兴”，那么最终将输出 4 个语义标签：“主语”、“状态”、“动词”、“高兴”。

在以上过程中，通过分析给定的文本，能够提取到文本在语义层面的特征信息，并将这些特征信息以语义标签这种简练的方式来进行表示，方便了动作生成过程中以语义层面的语义标签来作为指导信号，进而有利于合成与音频的语义高度匹配、圆融自然的虚拟形象肢体动作。

203、服务器从预设动作库中，检索与该语义标签相匹配的动作类别和属于该动作类别的动作数据，该预设动作库包括该虚拟形象的、属于多种动作类别的动作数据。

在一些实施例中，针对步骤 202 中获取到的每个语义标签，以该语义标签作为索引，从预设动作库的多个候选类别中，检索到与该语义标签相匹配的动作类别，其中，预设动作库是服务器侧创建并维护的一个动作数据库，用于以动作类别为单位，存放每个动作类别的动作集合，每个动作集合中包含聚类到这一动作类别的动作数据，预设动作库的创建方法将在后续实施例中详细说明，此处不再赘述。

但需要说明的是，在一种可能的实施情况下，并不是每个语义标签都能找到一个匹配的

动作类别的，如果语义标签与所有的候选类别都不匹配，那么可以将一个预设动作类别作为与该语义标签相匹配的动作类别，以避免动作序列中存在一段时间的空缺。其中，预设动作类别可以是技术人员预先配置的一种默认动作类别，比如没有语义的站立动作类别，或者静坐动作类别等，这里对预设动作类别不进行具体限定，技术人员对于不同的虚拟形象还可以配置不同的预设动作类别。

在以上过程中，能够以文本的语义标签作为索引，在预设动作库中检索与音频在语义层面最匹配的动作类别，这个动作类别并不是简单随着音频节奏而进行律动，而是能够与音频的语义信息高度适配，能够反映出来虚拟形象在播报音频的情感倾向和潜在语义，这样从该动作类别中挑选出来的动作数据，能够为虚拟形象合成准确率更高的动作序列。

在检索到与音频相匹配的动作类别后，从预设动作库中检索属于该动作类别的动作数据，该动作数据可以控制虚拟形象呈现出某种特定动作。

204、服务器基于该动作数据，生成该虚拟形象的动作序列，该动作序列用于控制该虚拟形象执行配合该音频的动作。

在一些实施例中，对每个语义标签配置了动作类别以后，对每个语义标签可以从预设动作库中检索到属于该动作类别的动作数据，例如，该动作数据可以包含连续时刻下的多帧 3D 骨骼数据（即每帧 3D 骨骼数据可以称为一个动作帧），每帧 3D 骨骼数据至少包含这一帧呈现的动作画面中每个骨骼关键点的位姿数据，这样，只需要将每帧 3D 骨骼数据迁移到虚拟形象的 3D 骨架模型，就能够控制虚拟形象呈现出某种特定动作。接着，可以按照每个语义标签对应的词语在音频中的时间戳顺序，将每个语义标签匹配到的动作数据进行拼接，形成虚拟形象的动作序列，这个动作序列表征了虚拟形象在播报音频的连续时刻下的肢体动作变化情况，用于控制虚拟形象在播报音频时执行配合该音频的肢体动作。

在一些实施例中，根据音素对齐工具，能够在音频时间轴上找到每个语义标签对应的一段时间戳区间，这个时间戳区间是指虚拟形象在播报属于这个语义标签的词语时的时间段，接着，从预设动作库中该动作类别的动作集合中，检索到与该语义标签相匹配的动作数据，进而使用该动作数据来填充动作序列中的这一时间戳区间，每个首尾相连的时间戳区间内的动作数据将构成虚拟形象在连续时刻下的动作序列。关于音素对齐方式，以及查询动作数据的方式，均会在下一实施例中详细说明，此处不再赘述。

在以上过程中，最终合成的动作序列中每一个动作帧都与音频中的一个音频帧时间戳对齐，使得动作帧反映出来音频帧在语义层面上相配合的肢体动作，使得声画适配度、准确度极大提升，不会产生机械死板的视觉效果，能够提升虚拟形象的仿真度、拟人度，优化虚拟形象的渲染效果。

需要说明的是，本申请实施例并不限定控制虚拟形象播报该音频以及执行配合该音频的动作的设备和时机。在一些实施例中，由服务器控制虚拟形象播报该音频以及基于该动作序列执行配合该音频的动作，在另一些实施例中，服务器将所生成的动作序列发送给关联的终端，由终端控制虚拟形象播报该音频以及基于该动作序列执行配合该音频的动作。并且，服务器生成动作序列后，可以立即控制虚拟形象播报该音频并基于该动作序列执行配合该音频的动作，或者，先将动作序列与音频或文本关联存储，后续在接收到播报指令的情况下再控制虚拟形象播报该音频并基于该动作序列执行配合该音频的动作。

上述所有可选技术方案，能够采用任意结合形成本公开的可选实施例，在此不再一一赘述。

本申请实施例提供的方法，通过以音频和文本作为双模态的驱动信号，在文本的基础上提取语义层面的语义标签，方便在预设动作库中检索到与语义标签匹配的动作类别，这个动作类别能够与音频的语义信息高度适配，反映出来虚拟形象在播报音频的情感倾向和潜在语义，进而检索属于该动作类别的动作数据，基于动作数据，为虚拟形象快速、高效地合成准确率更高的动作序列，不但提升虚拟形象的动作生成效率，且提升动作生成准确率。

进一步的，动作序列能够控制虚拟形象做出与音频在语义层面上配合的肢体动作，并非是简单跟随音频节奏进行律动，使得声画适配度、准确度极大提升，不会产生机械死板的视觉效果，能够提升虚拟形象的仿真度、拟人度，优化虚拟形象的渲染效果。

以上实施例中，简单介绍了虚拟形象的动作生成方案的流程，提出了一种音频和文本触发的肢体动作生成框架，由于虚拟形象在播报文本时，会发出音频，并表演肢体动作，因此音频、文本、肢体动作三者之间存在潜在的映射关系，并且能在音频时间轴对齐，本申请实施例中挖掘这种映射关系，在获取到音频和其文本以后，利用文本的语义标签，从预设动作库中检索语义层面与音频匹配的动作类别，进而根据属于该动作类别的动作数据，来合成虚拟形象的动作序列。以上动作生成方案，可适用于任意虚拟形象的肢体动作生成场景，比如，游戏人物、虚拟主播、影视人物、动漫人物等。

在本实施例中，将对虚拟形象的动作生成方案中每个步骤的具体实施方式进行详细说明。图 3 是本申请实施例提供的一种虚拟形象的动作生成方法的流程图。参见图 3，该实施例由计算机设备执行，以计算机设备为服务器为例进行说明，服务器可以为上述实施环境的服务器 102，该实施例包括以下步骤。

301、服务器获取虚拟形象的音频和文本，该文本指示该音频的语义信息。

其中，音频中包含至少一个音频帧，文本则是指示该音频的语义信息的文本，文本中包含至少一个词语，每个词语包含至少一个字符。音频和文本具有关联关系，即文本是对音频进行 ASR 识别到的语义信息，或者音频是播报该文本发出的语音信号，语音信号可以是机器输出的合成信号，也可以是麦克风采集的人声信号，这里对语音信号的类型不进行具体限定。

在一些实施例中，服务器从本地数据库中查询到一对具有关联关系的音频和文本，或者，服务器从本地数据库中取出一段音频，对该音频进行 ASR 识别，得到指示该音频的语义信息的文本，又或者，服务器从本地数据库中取出一段文本，对该文本进行声音合成，得到为该文本配音的音频。

在另一些实施例中，服务器从云端数据库中下载一对具有关联关系的音频和文本，或者，服务器从云端数据库中下载一段音频，对该音频进行 ASR 识别，得到指示该音频的语义信息的文本，又或者，服务器从云端数据库中下载一段文本，对该文本进行声音合成，得到为该文本配音的音频。

在又一些实施例中，服务器接收终端上传的一对具有关联关系的音频和文本，例如，终端向服务器发送动作生成请求，服务器接收并解析该动作生成请求，得到音频和文本。或者，服务器接收终端上传的音频，对该音频进行 ASR 识别，得到指示该音频的语义信息的文本，例如，终端向服务器发送动作生成请求，服务器接收并解析该动作生成请求，得到音频，对该音频进行 ASR 识别，得到指示该音频的语义信息的文本。或者，服务器接收终端上传的文本，对该文本进行声音合成，得到为该文本配音的音频，例如，终端向服务器发送动作生成请求，服务器接收并解析该动作生成请求，得到文本，对该文本进行声音合成，得到为该文本配音的音频。

在以上过程中，由于文本和音频可以互相转换，因此用户可以仅给定音频，也可以仅给定文本，还可以同时给定音频和文本，除了用户指定以外，也可以从本地数据库读取或者从云端数据库下载，本申请实施例对该音频和该文本的来源不进行具体限定。

一个示例性场景中，如图 4 所示，图 4 是本申请实施例提供的一种虚拟形象的动作生成方法的原理图，用户在终端侧输入音频和文本，终端将输入的该音频和该文本上传到服务器，服务器获取到音频 41 和文本 42“我第一次直播！”，其中音频 41 是虚拟形象播报文本 42 的音频文件，音频 41 可以是任意形式的音频文件，例如 WAV 文件、MP3 文件、MP4 文件等。

服务器获取到音频和文本后，执行本申请实施例提供的方法生成动作序列，该动作序列与文本的语义信息匹配，则后续在控制该虚拟形象播报该音频的过程中，控制该虚拟对象执行该动作序列所指示的肢体动作，以使该虚拟对象执行的肢体动作的语义信息与所播报的音

频匹配。

302、服务器基于该文本，确定该文本的情感标签。

其中，该情感标签表征该文本所表达的情感信息，如高兴、失落、愤怒等，本申请实施例对情感标签的内容不进行具体限定。

在一些实施例中，服务器中预先存储有多个候选的情感标签，并对每个候选的情感标签配置有多个情感关键词，存储情感关键词和情感标签的映射关系，进而提供一种基于关键词匹配的情感分析方法，如果该文本中包含任一情感关键词，那么基于该映射关系，可以查询到该情感关键词所映射到的情感标签，将查询到的情感标签作为该文本的一个情感标签；当然，如果该文本中包含多个情感关键词，那么每个情感关键词所映射到的情感标签都会作为该文本的情感标签。需要说明的是，如果多个情感关键词映射到同一个情感标签，那么还需要对该文本的情感标签进行去重。

以上基于关键词匹配的情感分析方式，计算量较小、计算复杂度较低，情感分析的速度快、效率高。

在另一些实施例中，服务器中预先存储有多个候选的情感标签，并对每个候选的情感标签配置一个情感特征，接着，对整个文本提取文本特征，将该文本特征和每个候选的情感标签的情感特征计算特征相似度，将特征相似度最高的情感标签作为该文本的情感标签。

进一步的，考虑到有时候播报文本的时候需要平铺直叙、不含情感倾向，那么技术人员还可以预先配置一个特征相似阈值，如果存在所有候选的情感标签的特征相似度小于该特征相似阈值，此时并不会挑选特征相似度最高的情感标签，此时情感标签空缺，或者使用一个默认的情感标签“无情绪”作为该文本的情感标签。这样能够提升情感标签的识别准确度，保证了不会对无情绪的文本添加不合适的情感标签。

进一步的，考虑到有时候播报文本的时候情绪组成是较为复杂的，有可能会共存多种情绪，因此，在预先配置特征相似阈值的情况下，还可以将特征相似度大于该特征相似阈值的情感标签作为该文本的情感标签。这样能够进一步提升情感标签的识别准确度，针对多种情绪交杂的文本具有更好的表现能力。

以上基于特征相似度的情感分析方式确定的情感标签的数量可以是 0 个、1 个或者 1 个以上，这里对情感标签的数量不进行具体限定。通过特征空间的相似度来评判整个文本的情感倾向，这样相较于关键词匹配的方式来说，情感分析的准确度更高，因为有些文本很可能本身不包含任何情感关键词，但在整个文本的语义层面上表达出较为明显的情感倾向，这种情况能够通过比较特征相似度检测出来。

在又一些实施例中，服务器中预先训练一个情感分析模型，将该文本输入该情感分析模型，通过该情感分析模型来计算该文本与每个候选的情感标签之间的匹配概率，接着，情感分析模型会基于该文本与每个候选的情感标签之间的匹配概率，输出与该文本相匹配的一个或多个情感标签，此种情况下，需要在候选的情感标签中添加一个“无情绪”的情感标签，就能够涵盖无情绪情况下的识别准确度，同样，技术人员也可以预先配置一个概率阈值，这样可以选择输出匹配概率最高的一个情感标签，其中，概率阈值是一个大于或等于 0 且小于或等于 1 的数值。或者输出匹配概率大于概率阈值的所有情感标签，或者输出匹配概率从大到小的排序中前 N（N≥1）个情感标签，本申请实施例对此不进行具体限定。可选地，情感分析模型可以是分类模型、决策树、深度神经网络、卷积神经网络、多层感知机等，本申请实施例对此不进行具体限定。

以上基于情感分析模型的情感分析方式，借助机器学习方法来学习文本与情感标签之间的潜在映射关系，从而评判文本与每个候选的情感标签之间的匹配概率，能够提升情感分析准确率，本申请实施例对情感分析方式不进行具体限定。

需要说明的是，步骤 302 是可选步骤，如果语义标签中不考虑情感标签，那么无需对文本进行情感分析，本申请实施例对是否必须进行文本情感分析不进行具体限定。

一个示例性场景中，仍以图 4 为例进行说明，通过以上任一种情感分析方式，针对文本 42“我第一次直播!”进行情感分析，得到文本 42 的情感标签“高兴”，说明虚拟形象播报文本 42 时需要沉浸在高兴情绪中。

303、服务器基于该文本，确定该文本中包含的至少一个词语。

在一些实施例中，服务器对文本进行分词，得到该文本的词语列表，该词语列表用于记录该文本中包含的至少一个词语，每个词语中包含至少一个字符。

分词过程可以使用分词工具实现，按照文本的语种不同，可以使用不同的分词工具，例如针对中文文本，使用中文分词工具来进行分词，得到中文文本的词语列表，又例如针对英文文本，使用英文分词工具来进行分词，得到英文文本的词语列表，本申请实施例对文本的语种不进行具体限定，对分词工具的类型也不进行具体限定。

一个示例性场景中，仍以图 4 为例进行说明，针对文本 42“我第一次直播!”进行分词，得到词语列表{“我”，“第一次”，“直播!”}，其中，文本 42 中包含 3 个词语，第 1 个词语“我”包含 1 个字符，第 2 个词语“第一次”包含 3 个字符，第 3 个词语“直播!”包含 3 个字符。

304、服务器从词性表中查询每个词语所属的词性标签。

其中，该词性标签表征该文本中词语的词性信息，如主语、动词、状态等，本申请实施例对词性标签的内容不进行具体限定。

在一些实施例中，服务器中预先存储有一个词性表，词性表中记录候选的词性标签，接着，对文本分词得到的每个词语，查询词性表，计算该词语的词向量与每个词性标签的标签向量之间的向量相似度，将向量相似度最高的词性标签作为该词语所属的词性标签。

一个示例性场景中，仍以图 4 为例进行说明，针对文本 42“我第一次直播!”进行分词，得到词语列表{“我”，“第一次”，“直播!”}，接着，在词性表中查询到第 1 个词语“我”所属的词性标签为“主语”，第 2 个词语“第一次”所属的词性标签为“状态”，第 3 个词语“直播!”所属的词性标签为“动词”。

在以上步骤 303~304 中，提供了提取文本中每个词语的词性标签的一种可能实施方式，这种查询词性表的方式计算量小、计算复杂度低，词性分析速度快、效率高。当然，也可以训练一个词性分析模型，将文本输入该词性分析模型中，由词性分析模型来输出一系列词语及其所属的词性标签，这样词性分析准确率较高，本申请实施例对词性分析方式不进行具体限定。

在以上词性分析过程中，因为说话时通常针对不同词性的词语会做出不同的肢体动作，因此词性的不同也会影响到肢体动作的类别或者幅度，考虑每个词语的词性标签，能够更好地反映出来文本在语义层面的隐含信息。

需要说明的是，步骤 304 是可选步骤，如果语义标签中不考虑词性标签，那么无需对文本进行词性分析（但还是需要进行分词的，因为只有分词以后才方便进行将词语、音素、动作进行三者对齐），本申请实施例对是否必须进行文本词性分析不进行具体限定。

305、服务器将该情感标签和该至少一个词语所属的词性标签，确定为该文本的语义标签。

其中，该语义标签表征该文本中词语的词性信息或者该文本表达的情感信息。

在一些实施例中，将步骤 302 获取到的情感标签和步骤 304 获取到的词性标签，确定为该文本的语义标签，其中，该语义标签的数量可以是一个或多个，本申请实施例对语义标签的数量不进行具体限定。

一个示例性场景中，仍以图 4 为例进行说明，针对文本 42“我第一次直播!”进行分词，得到词语列表{“我”，“第一次”，“直播!”}，其中，在词性表中查询到第 1 个词语“我”所属的词性标签为“主语”，第 2 个词语“第一次”所属的词性标签为“状态”，第 3 个词语“直播!”所属的词性标签为“动词”，另外，对文本 42 进行情感分析，得到该文本 42 所属的情感标签“高兴”，那么最终将输出 4 个语义标签：“主语”、“状态”、“动词”、“高兴”。以上分

析文本 42、提取语义标签的过程，称为“音频文本分析”过程。

在步骤 302~305 中，以语义标签既考虑词性标签又考虑情感标签为例，介绍了服务器确定该文本的语义标签的一种可能实施方式。通过分析给定的文本，能够提取到文本在语义层面的特征信息，并将这些特征信息以语义标签这种简练的方式来进行表示，方便了动作生成过程中以语义层面的语义标签来作为指导信号，进而有利于合成高度语义匹配、圆融自然的虚拟形象肢体动作。

需要说明的是，该语义标签可以包括词性标签或者情感标签中至少一项即可，如果语义标签不考虑词性标签，那么无需执行步骤 304，如果语义标签不考虑情感标签，无需执行步骤 302，本申请实施例对语义标签的内容不进行具体限定。

306、服务器对该文本中包含的每个词语，基于该词语关联的音素，从该音频中确定该音素所属的音频片段。

在一些实施例中，针对步骤 303 中文本中包含的每个词语，可以确定该词语关联的音素，其中，词语关联的音素指代播报该词语所需要发声的音素。每个词语关联的音素可以是一个或者多个，本申请实施例对音素的数量不进行具体限定。接着，从音频中找到该音素对应的至少一个音频帧，这至少一个音频帧就构成了该音素所属的音频片段。这样，能够将每个词语通过音素对齐方式，在音频中找到一个音频片段，从而将词语对齐到音频时间轴上的音频片段。

在一个示例性场景中，仍以图 4 为例说明，在步骤 303 中分词得到文本 42 中的每个词语以后，就可以进行音素对齐，即，确定播报该词语的 N（N≥1）个音素，在音频 41 中找到发出这 N 个音素的至少一个音频帧（例如第 2 帧到第 37 帧），将该至少一个音频帧作为该词语对齐的音频片段。以上过程，可视为对文本 42 的每个词语，从音频中确定一个对齐的音频片段的过程。

需要说明的是，本步骤 306 只需要在步骤 303 分词完毕以后就能够执行，与步骤 302 中提取情感标签、步骤 304 中提取词性标签均可以并行执行或者串行执行，本申请实施例不限制步骤 302、304 和 306 之间的执行时序。

307、服务器基于该词语所属的语义标签，从预设动作库中检索与该语义标签相匹配的动作类别和属于该动作类别的动作数据。

其中，该预设动作库包括该虚拟形象的、属于多种动作类别的动作数据。

在一些实施例中，步骤 305 中获取到的每个语义标签关联于文本中一个词语，对于语义标签中的词性标签来说，词性标签本身就是以词语为单位查询词性表得到的，因此词性标签和词语之间自然具有关联关系，每个词语必然属于一个词性标签，但不同词语有可能具有相同的词性标签；但对于语义标签中的情感标签来说，由于情感分析是以整段文本来进行分析的，这样综合整个文本的语境能够更好的判断出来其情感倾向，但也需要对情感标签在文本中找到一个最匹配的词语，例如，如果采用关键词匹配的情感分析方式确定情感标签，那直接将匹配到的情感关键词（必定是文本中的一个词语）作为情感标签最匹配的词语，如果采用基于特征相似度或情感分析模型的情感分析方式，那么在已知情感标签和每个分词得到的词语的情况下，反过来计算情感标签的词向量与每个词语的词向量之间的向量相似度，将向量相似度最高的词语作为情感标签最匹配的词语。

通过以上方式，不管语义标签中涵盖的是词性标签还是情感标签，都能够为每个语义标签找到一个最匹配的词语，需要说明的是，同一个词语可能会具有一个或多个语义标签，比如，仍以图 4 为例进行说明，文本 42“我第一次直播!”中，词语“直播!”具有 2 个语义标签，其中一个语义标签是词性标签“动词”，另一个语义标签则是情感标签“高兴”，本申请实施例对每个词语具有的语义标签的数量和类型都不进行具体限定。

这样，对文本中的每个词语来说，确定该词语所属的一个或多个语义标签以后，以该词语所属的每个语义标签作为索引，从预设动作库的多个候选类别中，查询得到与该语义标签

相匹配的动作类别，从而能够查询得到属于该动作类别的动作数据。

下面将以步骤 A1~A4 为例，介绍一种基于语义标签查询动作类别的可能实施方式，在这种实施方式中，从特征空间来评判语义标签是否与候选类别相似。

A1、服务器提取每个语义标签的语义特征。

在一些实施例中，对文本中每个词语所属的每个语义标签，服务器提取该语义标签的语义特征，例如，直接将该语义标签的词向量作为该语义特征，或者，预先训练一个特征提取模型，将该语义标签输入到该特征提取模型中，通过该特征提取模型对该语义标签进行处理，输出该语义标签的语义特征，该特征提取模型可以是任一种 NLP 模型，更进一步，为了提升特征提取效率，可以预先提取候选的全部词性标签和全部情感标签各自的语义特征，并将每个词性标签或情感标签和其自身的语义特征进行关联存储，这样，对每个语义标签，根据语义标签的标签 ID（Identification，标识），直接快速查询到与该标签 ID 关联存储的语义特征，这样相当于离线计算每个语义标签的语义特征，在线动作生成阶段只需要花费少量查询开销，不需要实时计算语义特征，能够提升特征提取效率。

在一些实施例中，以 Key-Value（键值对）数据结构来存储标签 ID 和其语义特征，其中，标签 ID 为 Key（键名），语义特征为 Value（键值），在线查询阶段中，以标签 ID 为索引，查询是否能够命中任一 Key-Value 数据结构，如果能够命中某个 Key-Value 数据结构，取出 Value 中存放的语义特征，这一语义特征就是标签 ID 所指示语义标签的语义特征。

A2、服务器查询该预设动作库中多个候选类别的类别特征。

在一些实施例中，服务器中创建和维护一个预设动作库，该预设动作库包括该虚拟形象的、属于多种动作类别的动作数据，动作库的构建流程将在下一实施例详细介绍，此处不再赘述。在预设动作库中储备有海量的动作数据，为了方便检索，将这些动作数据按照语义层级来进行聚类，从而划分成了多个动作类别，每个动作类别下有一个动作集合，该动作集合中存储被聚类到对应动作类别下的动作数据。在一些实施例中，该动作数据可以被实施为执行这一动作类别的动作时虚拟形象在连续时刻下的多帧 3D 骨骼数据。

进一步的，预设动作库中的全部动作类别构成了当前语义标签的多个候选类别，此时服务器可以对每个候选类别计算类别特征，例如，将候选类别的词向量作为该候选类别的类别特征，又例如，复用步骤 A1 中使用的特征提取模型，将候选类别输入到特征提取模型中，通过该特征提取模型对该候选类别进行处理，输出该候选类别的类别特征，这里仅以复用步骤 A1 中的特征提取模型为例进行说明，能够节约服务器侧的训练开销，无需重新训练一个特征提取模型，而且能够将语义标签和动作类别投影到同一个特征空间中，当然，服务器侧也可以对语义标签训练一个语义特征提取模型，对动作类别训练一个类别特征提取模型，使得语义特征和类别特征的提取过程更有针对性，分别提升语义特征和类别特征的表达能力，本申请实施例对此不进行具体限定。

更进一步，为了提升特征提取效率，可以预先利用训练完毕的特征提取模型，提取预设动作库中全部动作类别（即全部候选类别）的类别特征，接着，将每个动作类别和其自身的类别特征进行关联存储。这样，在线动作生成阶段，对每个候选类别，根据候选类别的类别 ID，直接快速查询到与该类别 ID 关联存储的类别特征，这样相当于离线计算每个候选类别的类别特征，从而在线查询时只需要花费少量查询开销，不需要实时计算类别特征，能够提升特征提取效率。

在一些实施例中，以 Key-Value 数据结构来存储类别 ID 和其类别特征，其中，类别 ID 为 Key，类别特征为 Value，在线查询阶段中，以类别 ID 为索引，查询是否能够命中任一 Key-Value 数据结构，如果能够命中某个 Key-Value 数据结构，取出 Value 中存放的类别特征，这一类别特征就是类别 ID 所指示候选类别的类别特征。

需要说明的是，在本申请实施例中，为了区别语义标签匹配到的动作类别，和用于候选的动作类别，才区分候选类别和动作类别的称呼，即候选类别和动作类别是相对于语义标签

而言的，但对于预设动作库本身来说，所有类别都是预设动作库所支持的动作类别，而没有候选类别这个概念。

A3、服务器从该多个候选类别中确定该动作类别，该动作类别的类别特征与该语义特征符合相似条件。

其中，该相似条件表征该语义标签与该候选类别是否相似。

在一些实施例中，对文本中每个词语所属的每个语义标签，服务器从步骤 A1 中获取到该语义标签的语义特征，从步骤 A2 中获取到预设动作库中全部候选类别的类别特征。接着，计算该语义特征与每个候选类别的类别特征之间的特征相似度，并该多个候选类别中，挑选特征相似度符合该相似条件的候选类别作为与该语义标签相匹配的动作类别，也即是确定的动作类别的类别特征与语义特征符合相似条件。其中，该特征相似度可以是余弦相似度、欧氏距离的倒数等，本申请实施例对此不进行具体限定。

在一些实施例中，相似条件为特征相似度最高，那么只需要从全部候选类别中，找到特征相似度最高的候选类别作为与该语义标签相匹配的动作类别即可，这样能够保证每个语义标签都一定能找到一个语义层面最为相似的动作类别，不会出现某些语义标签匹配不到动作类别的情况，其动作类别筛选流程较为简单，计算效率高。

在另一些实施例中，相似条件为特征相似度大于预设相似阈值，预设相似阈值为技术人员预先定义的大于 0 的数值，如果满足相似条件的候选类别只有一个，那么将唯一一个候选类别作为与该语义标签相匹配的动作类别，如果满足相似条件的候选类别多于一个，那么挑选特征相似度最大的候选类别作为与该语义标签相匹配的动作类别，如果满足相似条件的候选类别为 0 个，即全部候选类别都不满足相似条件，进入步骤 A4。这样通过配置一个预设相似阈值，能够将某些播报时情绪较为平稳、不含有特定明显语义的情况考虑进入，这种情况下虚拟形象是较为平静的播报内容，并不需要做出具有某种语义的肢体动作（如果做出的话可能会显得比较夸张），而这种情况下其实每个特征相似度整体取值较低，如果不配置预设相似阈值，那么直接选取相对取值最大的特征相似度即可，如果配置预设相似阈值，那么将会提供一种全部候选类别都不匹配的策略，此时进入步骤 A4，直接将与该语义标签相匹配的动作类别，配置成不具有特殊语义的预设动作类别，如站立动作类别、静坐动作类别等。

在以上步骤 A1~A3 中，提供了一种从特征空间来评判语义标签是否与候选类别相似的实施方式，这样能够对任一语义标签，找到在特征空间中与该语义标签满足相似条件的动作类别，通过控制相似条件，即可灵活控制是否使用预设动作类别来填充语义标签和全部候选类别均不太相似的情况，因此提升动作类别的识别效率，提升动作类别的可控性。

A4、在该多个候选类别的类别特征与该语义特征均不符合该相似条件的情况下，服务器将与该语义标签相匹配的动作类别配置为预设动作类别。

在一些实施例中，通过相似条件，并不一定每个语义标签都能找到相匹配的动作类别，如果语义标签的语义特征与所有的候选类别的类别特征都不符合该相似条件，说明语义标签和所有候选类别都不匹配，那么可以将一个预设动作类别作为与该语义标签相匹配的动作类别，以避免动作序列中存在一段时间的空缺。其中，预设动作类别可以是技术人员预先配置的一种默认动作类别，比如没有语义的站立动作类别，或者静坐动作类别等，这里对预设动作类别不进行具体限定，技术人员对于不同的虚拟形象还可以配置不同的预设动作类别。

在以上步骤 A1~A4 中，提供了以语义标签为单位来选取每个语义标签的动作类别的一种可能实施方式，能够以文本的语义标签作为索引，在预设动作库中找到与音频在语义层面最匹配的动作类别，这个动作类别并不是简单随着音频节奏而进行律动，而是能够与音频的语义信息高度适配的，能够反映出来虚拟形象在播报音频的情感倾向和潜在语义，这样从该动作类别中挑选出来的动作数据，能够为虚拟形象合成准确率更高的动作序列。

在另一些实施例中，除了步骤 A1~A4 以外，还可以训练一个动作分类模型，将每个语义标签输入到该动作分类模型中，通过该动作分类模型来预测语义标签与每个候选类别的匹配

概率，并输出匹配概率最高的动作类别，这样的话，只需要在候选类别中添加上述预设动作类别，也能够涵盖到播报时不需要做出包含语义的肢体动作的场景，能够进一步提升动作类别的识别准确度。

需要说明的是，由于每个语义标签都会关联于一个词语，但每个词语却可能具有多个语义标签，为了使得词语和动作类别一一对应，针对一个词语具有多个语义标签的情况，有可能会存在多个匹配的动作类别，此时，找到每个语义标签相匹配的动作类别以后，优先选择同时与该词语的全部语义标签相匹配的动作类别，作为这个词语最终挑选的动作类别，如果不存在同时与该词语的全部语义标签相匹配的动作类别，那么优先选择特征相似度更高的动作类别，或者直接配置成预设动作类别。举一个例子，假设某个词语具有 2 个语义标签 a 和 b，语义标签 a 匹配到了动作类别 1、2，语义标签 b 匹配到了动作类别 1、3，那么直接选中动作类别 1 作为这个词语最终的动作类别，但如果语义标签 b 匹配到了动作类别 3、4，那么将从动作类别 1~4 中选取特征相似度最高的动作类别，或者直接将预设动作类别作为这个词语最终的动作类别。

在一个示例性场景中，仍以图 4 为例，将音频文本分析阶段得到的每个语义标签作为索引，从预设动作库 43 的 K（K≥2）个动作类别中，挑选出与该语义标签相匹配的动作类别，例如，文本 42 中的 3 个词语“我”、“第一次”、“直播!”，第 1 个词语“我”只有 1 个语义标签“主语”，但语义标签“主语”在预设动作库 43 中没有找到匹配的动作类别，因此配置为预设动作类别“站立”，第 2 个词语“第一次”只有 1 个语义标签“状态”，但语义标签“状态”在预设动作库 43 中找到了匹配的动作类别“卖萌耸肩”，第 3 个词语“直播!”则有 2 个语义标签“动词”（词性标签）和“高兴”（情绪标签），其中语义标签“动词”和“高兴”共同锁定了一个动作类别“高兴举手”，即动作类别“高兴举手”同时与 2 个语义标签“动词”和“高兴”均匹配，所以被选中为第 3 个词语“直播!”最匹配的动作类别。

其中，预设动作库 43 也称为包含海量动作数据的动态语义预设动作库，海量动作数据可以是收集的、公开的、合规的虚拟形象 3D 动作片段的数据，例如每个 3D 动作片段包含在连续时刻下的多帧 3D 骨骼数据。以上根据语义标签检索动作类别的过程，也称为检索每个语义标签的关键 Pose（检索关键动作）。

进一步地，由于预设动作库 43 的数据量级有可能很大，在每个动作类别之中还可以继续划分成多个子类，比如动作类别“举手”中还划分多个子类：“举单手”、“举双手”等，在一个示例中，如图 4 所示，动作类别 1 包含 10 个子类，动作类别 2 包含 3 个子类，动作类别 3 包含 6 个子类……以此类推，动作类别 K 包含 2 个子类，这里对每个动作类别是否划分子类不进行具体限定。

需要说明的是，对每个语义标签来说，在每个动作类别中包含子类的情况下，还可以通过步骤 A1~A3 同理的方式，通过计算特征相似度，从已经确定的动作类别的全部子类中找到与该语义标签相匹配的子类，能够进一步提升步骤 308 中使用的动作数据与语义标签之间在语义层面的匹配程度。

308、服务器基于该词语对应的动作数据，生成与该音频片段相匹配的动作片段。

在上述步骤 307 中，对每个词语都能够找到一个唯一对应的动作类别，分为以下几种情况进行总结：1）该词语具有一个语义标签，如果该语义标签有满足相似条件的动作类别，那么选择这一动作类别，如果该语义标签没有满足相似条件的动作类别，那么选择预设动作类别；2）该词语具有多个语义标签，每个语义标签都按照步骤 1）选择了动作类别（含预设动作类别）以后，如果存在某个动作类别同时与该词语的全部语义标签相匹配，那么选择该同时匹配的动作类别，如果同时匹配的动作类别也有多个，那么选择同时匹配且特征相似度最高的动作类别，如果不存在动作类别同时与该词语的全部语义标签相匹配，那么选择匹配的语义标签数量最多的动作类别，或者选择特征相似度最高的动作类别，或者选择预设动作类别，本申请实施例对此不进行具体限定。

在上述基础上，每个词语具有一一对应的动作类别（含预设动作类别），那么根据音频时间轴的对应关系，每个词语在步骤 306 中能够找到一个音频片段，在步骤 307 中能够找到一个动作类别，根据预设动作库中属于该动作类别的动作数据，可以为该词语合成一个动作片段，从而保证动作片段和音频片段的时间戳对齐，且语义层面高度适配。

下面，将通过步骤 B1~B2 介绍一种可能的动作片段合成方式，在这种合成方式中，能够实现音频帧和关键动作帧的一一对应，使得两者时间戳对齐。

B1、服务器从该动作数据中，确定与该词语的语义匹配度最高的至少一个关键动作帧。

在一些实施例中，由于词语和动作类别是一一对应的关系，那么针对每个词语，服务器从预设动作库中检索到属于该词语所对应动作类别的动作数据，进而对动作数据进行筛选，得到与该词语的语义匹配度最高的至少一个关键动作帧。

在一些实施例中，预设动作库中每个动作类别存储一个动作集合，该动作集合用于存放属于该动作类别的动作数据，例如，该动作集合包含多个动作片段，每个动作片段都包含多个动作帧，每个动作帧表示虚拟形象在执行该动作类别下某个动作的过程中某一时刻下每个骨骼关键点的位姿，其中，每个动作片段具有其标注的参考音频和参考文本。由于在建库阶段，本身也会对参考文本中的词语、参考音频中的音素以及动作片段中的动作帧，三者进行时间戳对齐。因此，在比较词语和关键动作帧的语义匹配度的时候，可以先查询动作集合中是否有动作片段的参考文本中含有该词语，如果查询到了命中某个参考文本，那么直接从命中的参考文本对应的动作片段中，取出与该词语匹配（即时间戳对齐）的至少一个关键动作帧；如果没有查询到命中任何参考文本，那么需要进一步计算当前词语的词向量与每个参考文本中的每个词语的词向量之间的向量相似度，找到向量相似度最高的近似词语（通常是同义词和/或近义词）所属的参考文本，并从找到的参考文本对应的动作片段中，取出该近似词语匹配（即时间戳对齐）的至少一个关键动作帧。

在上述过程中，提供了一种从动作集合的动作数据中筛选关键动作帧的可能实施方式，这样先检测重复词语再检测近似词语的方式，能够保证只有在找不到重复词语的情况下，才需要检测近似词语，从而降低服务器的计算开销。在另一些实施例中，在检测近似词语时，也可以不基于向量相似度来评判近似词语，而是直接先从词表中为该词语获取同义词和/或近义词，再以同义词和/或近义词作为索引来查询是否命中某个参考文本，这样同样满足在找不到重复词语的情况下，才需要查询同义词和/或近义词，同样能够实现节约计算开销的效果。

在又一些实施例中，还考察一种情况，如果预设动作库中每个动作类别之下还细分了多个子类，那么在步骤 307 的可选方式中，能够在该动作类别的多个子类中找到与该词语相匹配的子类，这样在检索关键动作帧的阶段，只需要考虑属于选中的子类的动作数据，而不需要考虑属于未选中的子类的动作数据，相当于降低了关键动作帧的查询范围，进一步地提升了关键动作帧的查询效率，而且通常小范围查询到的关键动作帧，由于动作类别（即大类）和所属子类（即小类）双重匹配，从而也提升了关键动作帧的检索精度，能够更好地与该词语进行语义层面的协调和配合。

在另一些实施例中，从预设动作库中对每个动作类别仅保存一个标准的动作片段，那么只需要从动作片段最中间的中值动作帧开始，采样中值动作帧最邻近的至少一个关键动作帧即可，这样能够将关键动作帧对齐到预存的标准的动作片段的中间，往往位于中间的是比较标准和关键的动作/姿态。

B2、服务器基于该音频片段，将至少一个关键动作帧合成为与该音频片段相匹配的动作片段。

在一些实施例中，对文本中的每个词语，服务器基于步骤 B1 查询到至少一个关键动作帧以后，可以确定该关键动作帧的帧数，此外，还确定步骤 306 中与该词语的时间戳对齐的音频片段的音频帧数，进而比较该音频帧数和该关键动作帧的帧数，可选地，根据音频帧数，对关键动作帧进行一定比例的倍速缩放，保证最终合成的动作片段与步骤 306 的音频片段等

长(即时间戳对齐),这种情况下不需要对关键动作帧进行裁剪或修饰,只需要调整播放倍速,因此总体来说能够保留关键动作帧的较多细节,尽量呈现词语所匹配到关键动作的完整姿态变化。

在另一些实施例中,如果该音频帧数和该关键动作帧的帧数两者差距较为悬殊,单纯调整倍速有可能会导致播放衔接不流畅,如虚拟形象突然动作迟缓,或者虚拟形象突然动作飞快,这样显然会影响动作流畅度和自然度。因此,在本申请实施例还提供一种对关键动作帧进行插帧或者裁剪的方式,来改善上述涉及的情况,优化动作流畅度和自然度。下面,将通过两种情况来分类讨论,分别涉及到关键动作帧的帧数不超过音频片段的音频帧数的情况一,以及关键动作帧的帧数超过音频片段的音频帧数的情况二。

情况一、关键动作帧的帧数不超过音频片段的音频帧数

在一些实施例中,在该关键动作帧的帧数不超过该音频片段的音频帧数的情况下,服务器可以对该至少一个关键动作帧进行插帧,得到与该音频片段等长的该动作片段。

在一些实施例中,为了保证动作片段与音频片段等长,可以对该至少一个关键动作帧进行插帧,比如,在任一对或者多对相邻的关键动作帧之间插入一个或多个中间动作帧,每个中间动作帧是根据其插入的那一对相邻的关键动作帧所计算出来的中间动作数据。

在一些实施例中,采用线性插帧方式,那么计算中间动作数据,实际上采用线性插值法来进行计算,比如,在关键动作帧1和关键动作帧2中插入i(i≥1)个中间动作帧,以左肩的同一骨骼关键点为例,该骨骼关键点在关键动作帧1中处于位姿θ_1,该骨骼关键点在关键动作帧2中处于位姿θ_2,那么只需要计算该骨骼关键点从位姿θ_1变换至位姿θ_2的i个中间位姿,就能够得到i个中间动作帧中这一骨骼关键点的i个中间位姿,以此类推,对全身的骨骼关键点计算i个中间位姿,就能够实现插入i个中间动作帧。在线性插值法中,骨骼关键点在i个中间动作帧中是按照固定步长均匀变化的,也可以认为骨骼关键点匀速运动,因此只需要根据初始状态和末尾状态下的位姿(即位姿θ_1和位姿θ_2),计算出来插入i个中间动作帧时的固定步长,就很容易实现i个中间位姿的计算。以上线性插帧方式,计算资源消耗少,计算开销低,动作片段合成快,等待延时低。

在另一些实施例中,预先训练一个动作调节模型,该动作调节模型用于在关键动作帧的帧数小于音频片段的音频帧数的情况下,对关键动作帧进行非线性插帧,即,该动作调节模型用于学习到关键动作帧的非线性插帧模式,这种非线性插帧模式可能是按照某种运动曲线来进行拟合的,也可能是按照音频节奏还拟合运动幅度从而学习到这种幅度变化下的位姿变化规律,具体学习哪种非线性插帧模式是由投入的训练样本来决定的,在对动作调节模型训练完毕以后,即可将至少一个关键动作帧输入该动作调节模型,并以音频帧数作为超参数来进行控制,进而动作调节模型会输出待插入的至少一个中间动作帧,并且相邻两个关键动作帧中插入的每个中间动作帧之间并非按照固定步长均匀变化,而是根据动作调节模型学习到的非线性插帧模式进行非匀速的位姿变化,本申请实施例对是否采用线性插帧方式不进行具体限定。以上基于动作调节模型的非线性插帧方式,能够改善线性插帧方式可能带来的机械感,优化动作片段的流畅度。

在上述过程中,针对关键动作帧的帧数少于音频帧数的情况下,通过对关键动作帧进行插帧,能够补足缺失的动作帧,使得相邻关键动作帧之间补充中间的运动状态,这样动作片段中虚拟形象会运动更加连贯。

情况二、关键动作帧的帧数超过音频片段的音频帧数

在一些实施例中,在该关键动作帧的帧数超过该音频帧数的情况下,创建与该音频片段等长的动作片段,将该动作片段的每一帧填充为预设动作类别下的预设动作帧。其中,预设动作类别可以是技术人员预先配置的一种默认动作类别,比如没有语义的站立动作类别,或者静坐动作类别等,这里对预设动作类别不进行具体限定,技术人员对于不同的虚拟形象还可以配置不同的预设动作类别。其中,预设动作帧是预设动作类别下预先配置的一种相对静

止的动作帧，比如预设动作类别是站立动作类别时，预设动作帧就是站立动作帧，预设动作类别是静坐动作类别时，预设动作帧就是静坐动作帧，而在保持预设动作类别时虚拟形象通常在多帧保持相同动作不变的。

在上述过程中，针对关键动作帧的帧数超过音频帧数的情况下，通过丢弃掉这部分关键动作帧，并使用预设动作帧来填充这个动作片段，这样既不需要高倍速播放关键动作帧导致观众体验受损，也不会导致确实某个动作片段出现问题。

在另一些实施例中，在该关键动作帧的帧数超过该音频帧数的情况下，还可以对该关键动作帧进行裁剪，比如在丢弃掉首尾的一部分关键动作帧，以使裁剪以后关键动作帧的帧数不超过音频帧数，这样避免使用预设动作帧填充某个较长的词语的音频片段，动作生成效果更好，但有可能会破坏关键动作帧的完整度，这种情况下需要通过步骤 309 中的动作平滑操作来进行改善，至于首尾关键动作帧的裁剪逻辑可以由技术人员进行配置，比如按照设定帧数进行裁剪，或者按照设定比例进行裁剪，本申请实施例对此不进行具体限定。

在步骤 B1~B2 中，提供了一种可能的动作片段合成方式，在这种合成方式中，能够实现音频帧和关键动作帧的一一对应，使得两者时间戳对齐。即使在关键动作帧的帧数和音频帧数不适配的时候，也可以通过插帧、裁剪或者填充预设动作帧的方式，保证动作片段顺利合成，提升动作片段合成效率。

309、服务器基于每个词语的音频片段相匹配的每个动作片段，生成与该音频相匹配的动作序列，该动作序列用于控制该虚拟形象执行配合该音频的动作。

在一些实施例中，针对文本中的每个词语，在步骤 306 中能够找到一个唯一对应的音频片段，在步骤 308 中又能够合成一个唯一对应的动作片段，因此步骤 306 的音频片段和步骤 308 的动作片段，能够以词语为桥梁实现三者的一一对应，且时间戳对齐。这样，只需要按照每个音频片段的时间戳顺序，将每个动作片段依次拼接，即可得到一个动作序列，且保证动作序列中每个动作片段都与音频中的一个音频片段在语义层面上高度适配。

在另一些实施例中，还可以对拼接得到的动作序列进行动作平滑，来增加不同动作片段衔接时的自然度和流畅度，下面将通过步骤 C1~C2 来进行详细说明。

C1、服务器基于每个音频片段的时间戳顺序，拼接每个音频片段相匹配的每个动作片段，得到拼接动作序列。

在一些实施例中，由于音频片段和动作片段以词语为桥梁实现三者一一对应，因此，对于每个动作片段来说，可以在音频时间轴上找到对应的音频片段的时间戳区间，进而按照时间戳区间的先后顺序，对每个动作片段进行拼接，得到一个拼接动作序列。可选地，直接输出拼接动作序列，简化动作合成流程，或者，执行步骤 C2 中的动作平滑操作，增加不同动作片段衔接时的自然度和流畅度。

C2、服务器对该拼接动作序列中的每个动作帧进行动作平滑，得到该动作序列。

在一些实施例中，由于步骤 C1 得到的拼接动作序列中，有些可能是关键动作帧，有些可能是插帧的中间动作帧，还有些可能是填充的预设动作帧，因此将拼接动作序列中的每一帧动作数据称为一个动作帧，动作帧可以是关键动作帧、中间动作帧或者预设动作帧，本申请实施例对此不进行具体限定。接着，对拼接动作序列中的每个动作帧进行动作平滑，得到最终的动作序列。

在一些实施例中，使用窗口平滑方式，对连接起来的每个动作帧进行全局处理，得到一个全局平滑以后的动作序列。其中，窗口平滑方式是指：以骨骼关键点为单位，确定同一骨骼关键点在每个动作帧中的位姿，这样能够得到骨骼关键点在动作序列中的一系列位姿变化，从而能够拟合出一条位姿变化折线，进而通过移动窗口平均平滑算法，对该位姿变化折线进行平滑，得到一条位姿变化曲线，进而再按照时间戳从位姿变化曲线中采样骨骼关键点在每个动作帧中的位姿，从而得到骨骼关键点在每个动作帧中的更新位姿。这样能够在相邻两个动作片段匹配到的动作类别差距较大时，通过窗口平滑方式使得相邻两个动作片段的衔接更

加流畅、连贯、自然，生成视觉效果更好的动作序列，提升动作合成的准确率。

在另一些实施例中，除了窗口平滑方式以外，还可以对位姿变化折线采用其他平滑算法进行平滑，或者直接在位姿变化折线上机器拟合出一条位姿变化曲线，这样同样能够达到动作平滑的效果。

在步骤 C1~C2 中，通过将机械拼接形成的拼接动作序列进行动作平滑，使得相邻两个动作片段的衔接更加流畅、连贯、自然，生成视觉效果更好的动作序列，提升动作合成的准确率。当然，也可以直接输出拼接动作序列而不进行动作平滑，这样简化动作合成流程，提升动作合成效率。

在一个示例性场景中，仍以图 4 为例进行说明，针对文本 42“我第一次直播!”，第 1 个词语“我”匹配到预设动作类别“站立”，第 2 个词语“第一次”匹配到动作类别“卖萌耸肩”，第 3 个词语“直播!”匹配到动作类别“高兴举手”。那么，将合成 3 个动作片段：站立动作片段、卖萌耸肩动作片段和高兴举手动作片段，动作片段合成方式详细参考步骤 308，此处不再赘述。接着，将站立动作片段、卖萌耸肩动作片段和高兴举手动作片段三者拼接，得到一个拼接动作序列，对拼接动作序列进行动作平滑，得到最终输出的动作序列。图 4 中还输出一条平滑以后的位姿变化曲线（即动作曲线），表征输出的动作序列中骨骼关键点的位姿变化曲线是较为平滑、流畅的，能够去除肢体动作的机械感。可选地，本申请实施例适用于合成虚拟形象的肢体动作，但还需要结合虚拟形象的面部表情，才能够生成最终的画面，将画面和音频结合才能够生成最终的虚拟形象视频（如数字人视频）。

在步骤 306~309 中，提供了基于动作数据，生成该虚拟形象的动作序列的一种可能实施方式，针对每个动作类别具有海量动作数据的情况下，都能够挑选出语义匹配度最高的、具有代表性的关键动作帧，进而合成一系列动作片段，再拼接成一个动作序列，这个动作序列表征了虚拟形象在播报音频的连续时刻下的肢体动作变化情况，用于控制虚拟形象在播报音频时执行配合该音频的肢体动作。

在以上过程中，最终合成的动作序列中每一个动作帧与音频中的一个音频帧的时间戳对齐，使得动作帧反映出来音频帧在语义层面上相配合的肢体动作，使得声画适配度、准确度极大提升，不会产生机械死板的视觉效果，能够提升虚拟形象的仿真度、拟人度，优化虚拟形象的渲染效果。

上述所有可选技术方案，能够采用任意结合形成本公开的可选实施例，在此不再一一赘述。

本申请实施例提供的方法，通过以音频和文本作为双模态的驱动信号，在文本的基础上提取语义层面的语义标签，方便在预设动作库中检索到与语义标签匹配的动作类别，这个动作类别能够与音频的语义信息高度适配，反映出来虚拟形象在播报音频的情感倾向和潜在语义，进而检索属于该动作类别的动作数据，基于动作数据，为虚拟形象快速、高效地合成准确率更高的动作序列，不但提升虚拟形象的动作生成效率，且提升动作生成准确率。

并且，动作序列能够控制虚拟形象做出与音频在语义层面上配合的肢体动作，并非是简单跟随音频节奏进行律动，使得声画适配度、准确度极大提升，不会产生机械死板的视觉效果，能够提升虚拟形象的仿真度、拟人度，优化虚拟形象的渲染效果。

在以上动作生成方案中，挖掘到了音频的文本和肢体动作之间潜在的映射关系，实现了文本和音频双模态触发虚拟形象肢体动作生成的自动化流程，无需人工干预，既不需要真人表演结合动作捕捉系统，也不需要动画师进行动画修复，能够由机器在给定文本和音频的情况下，快速、自动地生成虚拟形象其肢体动作的动作序列，替代了繁琐的动作捕捉和修复流程，且具有很强的通用性，能够使用到游戏、直播、动画、影视等各种场景中虚拟形象的肢体动作生成任务，具有很高的实用性，并且其设备、人力、时间成本大幅降低，应用简单快速、无依赖性，且动作序列的生成品质高、准确性高。

在以上每个实施例中，详细介绍了虚拟形象的动作生成方案，能够在无需人工干预的情

况下，快速、自动化地在音频和文本的双模态驱动信号下，合成一段语义层级高度匹配的动作序列。以上动作生成方案依赖于构建完毕的预设动作库，而在本申请实施例中将对该预设动作库的建库流程进行详细说明。

图 5 是本申请实施例提供的一种虚拟形象的动作库的构建方法的流程图。参见图 5，该实施例由计算机设备执行，以计算机设备为服务器为例进行说明，服务器可以为上述实施环境的服务器 102，该实施例包括以下步骤。

501、服务器获取每个样本形象的样本动作序列、参考音频和参考文本，该参考文本指示该参考音频的语义信息，该样本动作序列用于控制该样本形象执行配合该参考音频的动作。

其中，样本形象是公开的、可收集的且合规收集到的虚拟形象或者真实形象，例如样本形象是动漫人物、虚拟主播、数字人等虚拟形象，也可以是演员、演讲者、主播等真实形象，本申请实施例对此不进行具体限定。

其中，样本形象的样本动作序列、参考音频和参考文本的收集和使用均是符合规定的，样本动作序列具有一一对应的参考音频（即配音）和参考文本（即字幕或者音频识别到的文本）。

在一些实施例中，服务器获取多个样本形象的样本动作序列，并剔除掉既没有标注参考音频也没有标注参考文本的低质量样本，进一步还可以剔除掉不包含肢体动作（比如视角只能看到虚拟形象的头部）的低质量样本，进一步还可以剔除掉持续时长太短或者太长的低质量样本，比如，仅保留持续时长位于 1~10s（秒）的样本动作序列，如果样本动作序列同时具有参考音频和参考文本，将三者对应存储，如果样本动作序列仅具有参考音频，那么对参考音频进行 ASR，得到对应的参考文本，将三者对应存储，如果样本动作序列仅具有参考文本，那么对参考文本进行配音（即基于文本进行语音合成），得到对应的参考音频，将三者对应存储。这里对样本形象的数量、样本动作序列的数量都不进行具体限定。

502、服务器基于该参考文本中词语和该参考音频中音素的关联关系，将该样本动作序列划分为多个样本动作片段，每个样本动作片段与该参考文本中的一个词语以及该参考音频中的一个音素相关联。

在一些实施例中，服务器以每个样本动作序列为单位进行处理，获取该样本动作序列对应存储的参考文本和参考音频，进一步的，通过上一实施例可知，通过音素对齐方式，能够建立该参考文本中词语和该参考音频中音素的关联关系，那么基于该关联关系，可以将该样本动作序列划分为多个样本动作片段。

在一些实施例中，通过下述步骤 D1~D2，介绍一种可能的样本动作片段的划分方式。

D1、服务器对该参考文本中的每个词语，基于该词语关联的音素，从该样本音频中确定该音素所关联的该样本音频片段。

上述步骤 D1 和上一实施例的步骤 306 同理，此处不再赘述。

D2、服务器基于每个样本音频片段的时间戳区间，将该样本动作序列划分成多个样本动作片段，每个样本动作片段与一个样本音频片段的时间戳区间对齐。

在一些实施例中，对每个样本音频片段，能够在音频时间轴上找到该样本音频片段中首个音频帧的开始时间戳和最后一个音频帧的结束时间戳，这个开始时间戳和结束时间戳构成了一个时间戳区间，由于参考音频、参考文本本身和样本动作序列就是时间戳对齐的，那么直接在样本动作序列中，按照每个样本音频片段的时间戳区间进行分割，既能够划分成多个样本动作片段，并保证每个样本动作片段的时间戳区间和样本音频片段的时间戳区间对齐。

503、服务器基于该样本动作片段的动作特征，对每个样本形象的每个样本动作片段进行聚类，得到多个动作集合，每个动作集合指示属于同一动作类别且属于不同样本形象的动作数据。

在一些实施例中，服务器对每个样本动作序列执行步骤 502 划分出多个样本动作片段，这样得到来自不同样本形象或者不同样本动作序列的一系列样本动作片段，接着，对每个样

本动作片段提取该样本动作片段的动作特征，可选地，训练一个动作特征提取模型，将样本动作片段输入到动作特征提取模型中，通过该动作特征提取模型对该样本动作特征进行处理，输出该样本动作片段的动作特征。

进一步的，在提取到每个样本动作片段的动作特征的基础上，基于聚类算法，对每个样本动作片段进行聚类，形成多个动作集合，每个动作集合代表一个动作类别，每个动作集合中包含属于相应的动作类别的动作数据(即聚类到这一动作类别的每个样本动作片段)，其中，聚类算法包括但不限于：KNN（K-Nearest Neighbor，K 近邻）聚类算法、K-means（K 均值）聚类算法、层次聚类算法等。

在一个示例性场景中，以 K 均值聚类算法为例，对样本动作片段的聚类过程进行说明，K 均值聚类算法是一种迭代求解的聚类分析算法，其步骤是：将全部样本动作片段分为 K 个动作类别，先随机选取 K 个样本动作片段作为 K 个动作类别各自初始的聚类中心，然后计算其余的每个样本动作片段与 K 个初始的聚类中心之间的距离（实际上计算的是动作特征之间的距离），把其余的每个样本动作片段分配给距离它最近的聚类中心，聚类中心以及分配给它们的其余的样本动作片段就代表一个动作集合。每向动作集合中新分配一个样本动作片段，该动作集合的聚类中心会根据已有的所有样本动作片段被重新计算。上述过程将不断重复直到满足某个终止条件，终止条件包括但不限于：没有（或最小数目）样本动作片段被重新分配给不同的动作集合，没有（或最小数目）聚类中心再发生变化，或者 K 均值聚类的误差平方和局部最小等，本申请实施例对 K 均值聚类算法的终止条件不进行具体限定。

在一个示例性场景中，如图 6 所示，图 6 是本申请实施例提供的一种动作库创建方法的原理图，以单个样本动作序列为例进行说明，获取样本动作序列的参考文本 61 和参考音频 62，例如参考文本 61 是“认识大家的第一天，开心”。接着，利用分词工具，将参考文本 61 分词成 5 个词语“认识”、“大家”、“的”、“第一天”、“开心”，接着，利用词性表查询到每个词语的词性标签，例如“认识”的词性标签为“v（动词）”，“第一天”的词性标签为“TIME（时间）”等，接着利用音素对齐工具，识别出每个词语对齐的样本音频片段的首帧序号和末帧序号，例如，“认识”的样本音频片段是 2~37 帧。通过以上操作，对每个词语能够构建一个四元组[词语，首帧序号，末帧序号，词性标签]，例如，词语“认识”的四元组为[‘认识’，2，37，‘v’]。将每个词语的四元组拼接，能够得到一个词性序列。接着，按照每个词语的样本音频片段，将样本动作序列划分成 4 个样本动作片段，其中由于词语“的”样本动作片段持续时间太短，将词语“大家”和“的”两者的样本动作片段合成了一个样本动作片段，将词语、样本音频片段和样本动作片段三者进行时间戳对齐。接着，对每个样本动作序列按照以上方式划分多个样本动作片段以后，将每个样本动作片段输入到聚类算法中，得到 K 个动作类别各自的动作集合，其中 K 为大于或等于 2 的整数。

在另一些实施例中，对每个动作类别的动作集合，还能够通过同理的方式来细分成多个子类，子类的聚类过程与动作类别的聚类过程同理，此处不再赘述，通过聚类的方式能够将海量的动作数据划分成多个动作类别，并保证每个动作类别内部的动作数据具有一定相似性，而不同动作类别之间的动作数据具有一定的差异性，这样就认为每个动作类别可以代表一个动作语义，即属于不同动作类别的动作数据在语义层面互不相同。

504、服务器基于该多个动作集合，构建动作库。

在一些实施例中，服务器基于步骤 504 中聚类形成的 K 个动作集合，直接构建动作库。动作库中包括 K 个动作集合，也即是包括虚拟形象的、属于 K 个动作类别的动作数据。

在一些实施例中，还计算并存储 K 个动作集合各自所属动作类别的类别特征。这样简化了动作库的创建流程，加快了动作库的建库效率。

在另一些实施例中，还可以对步骤 504 中聚类形成的 K 个动作集合进行进一步数据清洗，过滤掉每个动作集合中较为偏离聚类中心的离群样本，从而提升同一动作类别中每个样本动作片段的相似性，降低不同动作类别中每个样本动作片段的相似性。下面，以步骤 E1~E4 为

例，对单个动作集合的数据清洗流程进行说明。

E1、服务器对每个动作集合，获取该动作集合所指示的动作类别的类别特征，该类别特征为该动作集合中每个样本动作片段的平均动作特征。

在一些实施例中，对步骤 504 中聚类形成的每个动作集合，根据该动作集合中每个样本动作片段的动作特征，计算一个平均动作特征，作为该动作集合所指示的动作类别的类别特征，这一类别特征表征了该动作集合的聚类中心。

E2、服务器确定该动作集合中每个样本动作片段的动作特征对该类别特征的贡献度分数，该贡献度分数表征该样本动作片段与该动作类别的匹配程度。

其中，虽然每个样本动作片段归属于一个动作类别，但是不同的样本动作片段与动作类别的匹配程度可能不同。该匹配程度用于衡量该样本动作片段所做出的动作是否标准。例如，动作类别的类别特征为动作集合中的多个样本动作片段的平均动作特征，但是动作集合中有些样本动作片段的动作特征与该平均动作特征较相似，表示该样本动作片段做出的属于该动作类别的动作较为标准，而有些样本动作片段的动作特征与该平均动作特征不太相似，表示该样本动作片段虽然也做出了属于该动作类别的动作，但是所做的动作不够标准。因此该贡献度分数表征该样本动作片段相对于该动作类别的标准程度。

在一些实施例中，对该动作集合中的每个样本动作片段，计算该样本动作片段的动作特征对步骤 E1 中类别特征的贡献度分数，可选地，直接计算该动作特征和该类别特征之间的特征相似度，再对整个动作集合中每个样本动作片段的特征相似度进行指数归一化，得到每个样本动作片段的贡献度分数（指经过指数归一化以后的特征相似度）。这样通过指数归一化以后的特征相似度，来作为贡献度分数的度量指标，能够降低贡献度分数的计算复杂度，提升贡献度分数的计算效率。

在另一些实施例中，提供一种基于排除自身个体以后的类内方差（也称为 N-1 方差）来作为贡献度分数的度量指标，这样类内方差表征了被排除的个体对整个聚类的贡献度，也即体现了被排除的样本动作片段对整个动作集合的贡献度分数，其贡献度分数的表现能力更好，度量维度也更加精准，当贡献度分数越大时，说明样本动作片段的动作越标准，当贡献度分数越小时，说明样本动作片段的动作越不标准。下面，将通过步骤 E21~E22，对单个样本动作片段的类内方差（即一种可能的贡献度分数）的获取方式来进行详细说明。

E21、服务器对该动作集合中任一样本动作片段，获取除了该样本动作片段以外的每个其余动作片段的动作分数，该动作分数表征该其余动作片段与该类别特征的相似程度。

其中，其余动作片段是指动作集合中除了该样本动作片段以外的样本动作片段。

在一些实施例中，对该动作集合中的每个样本动作片段，计算该样本动作片段的动作特征与步骤 E1 中类别特征之间的特征相似度，再对整个动作集合中每个样本动作片段的特征相似度进行指数归一化，得到每个样本动作片段的动作分数（指经过指数归一化以后的特征相似度）。接着，排除当前的样本动作片段，确定除了该样本动作片段以外的每个其余动作片段的动作分数。

E22、服务器基于每个其余动作片段的动作分数，确定排除该样本动作片段以后的类内方差，将该类内方差确定为该样本动作片段的贡献度分数。

在一些实施例中，服务器计算步骤 E21 中获取到的全部其余动作片段的动作分数的平均值，将该平均值作为一个平均动作分数，再基于该平均动作分数和每个其余动作片段的动作分数，确定排除该样本动作片段以后的类内方差，将该类内方差确定为该样本动作片段的贡献度分数。

在一个示例中，假设动作集合中包含 N 个样本动作片段，以排除第 N 个样本动作片段为例，其余动作片段则是指从第 1 个至第 N-1 个样本动作片段，通过如下公式来获取上述类内方差（也称为 N-1 方差）：

$$S_{N-1}=\sqrt{\frac{\sum_{i}^{N-1}(x_i-\overline{x}_{N-1})}{N-1}}$$

其中，S_{N-1}表征第 N 个样本动作片段的类内方差，i 为大于或等于 1 且小于或等于 N-1 的整数，x_i表征第 i 个样本动作片段的动作分数，$\overline{x}_{N-1}$表征平均动作分数，其中平均动作分数是指这一共 N-1 个其余动作片段的动作分数的平均值。

在上述过程中，提供了一种基于排除自身个体以后的类内方差（也称为 N-1 方差）来作为贡献度分数的度量指标，这样实际上是排除指定的一个个体以后，计算其余个体的类内方差，那么类内方差越大，说明被排除的个体对偏离聚类的影响越小，其余个体对偏离聚类的影响越大，因此类内方差能够很好地衡量被排除的个体对整个聚类的贡献度，也即体现了被排除的样本动作片段对整个动作集合的贡献度分数，其贡献度分数的表现能力更好，度量维度也更加精准，当贡献度分数越大时，说明样本动作片段的动作越标准，当贡献度分数越小时，说明样本动作片段的动作越不标准，那么需要考虑剔除掉不标准的样本动作片段（即贡献度分数低的样本动作片段），这样方便进行每个动作类别内部的数据清洗。

E3、服务器从该动作集合中，剔除贡献度分数符合剔除条件的样本动作片段。

在一些实施例中，服务器可以按照贡献度分数从大到小的顺序，对该动作集合中的每个样本动作片段进行排序，剔除在该排序中位于末位的样本动作片段。这样每次数据清洗只会丢弃掉对偏离聚类的影响最小的样本动作片段，这样避免误删除掉高质量的样本动作片段。

在另一些实施例中，服务器还可以按照贡献度分数从大到小的顺序，对该动作集合中的每个样本动作片段进行排序，剔除在该排序中位于后 j 位的样本动作片段。这样每次数据清洗会丢弃掉对偏离聚类的影响较小的 j 个样本动作片段，这样通过灵活控制 j 的取值，就能够精细调控动作集合的数据清洗速率。其中，j 为大于或等于 1 的整数。

在一个示例中，如图 7 所示，图 7 是本申请实施例提供的一种动作集合的数据清洗原理图，针对某个动作类别的动作集合，将首个样本动作片段排除以后，计算剩下 N-1 个其余动作片段的类内方差，得到首个样本动作片段的贡献度分数为 0.2，对每个样本动作片段重复以上操作计算出来每个样本动作片段的贡献度分数，接着，按照贡献度分数从大到小的顺序，对每个样本动作片段进行排序，接着，剔除掉排序中位于末位的样本动作片段，例如剔除掉贡献度分数为 0.02 的末位的样本动作片段。

E4、服务器基于剔除后的动作集合，更新该类别特征和该贡献度分数，迭代多次执行剔除操作，在满足迭代停止条件的情况下停止迭代。

在一些实施例中，由于在步骤 E3 中剔除掉了一个（或多个）贡献度分数较低的样本动作片段，那么由于动作集合中的样本数量发生变化，其聚类中心即类别特征必然需要重新计算，因此基于步骤 E1 同理的方式更新该类别特征，相应地，由于类别特征发生变化，每个样本动作片段的贡献度分数也必然需要重新计算，因此基于步骤 E2 同理的方式更新该贡献度分数，再基于步骤 E3 同理的方式，按照更新后的贡献度分数，继续剔除掉贡献度分数符合剔除条件的样本动作片段。迭代执行步骤 E1~E3，直到满足迭代停止条件的情况下停止迭代，得到较为纯净的高质量的动作集合。其中，迭代停止条件包括但不限于：迭代次数到达次数阈值，次数阈值为大于 0 的整数；或，动作集合的样本容量缩减至预设容量，预设容量为大于或等于 1 的整数；或，排序位于末位的贡献度分数大于贡献度阈值，贡献度阈值是大于或等于 0 的数值，本申请实施例对迭代停止条件不进行具体限定。

在以上步骤 E1~E4 中，由于聚类直接形成的动作集合比较粗糙，有可能会存在一些类内差异较大的动作，这种动作需要被剔除，避免影响动作类别的聚类准确性，从而提供了一种对每个动作集合进行数据清洗、数据过滤或者说数据提纯的方式，最终基于清洗完毕的动作

集合构建的动作库，动作生成效果更好，可用性更高，整个迭代排序筛选的流程能够自监督实现，而不需要人工干预，因此建库阶段也能够自动化实现，建库成本低，建库效率高。

在上述步骤 501~504 中，详细介绍了为虚拟形象的动作生成方案提供支持的动作库的建库流程，在一些实施例中，考虑到动作库不可能一成不变，往往需要扩充或者新增一些动作数据。下面将步骤 F1~F4 为例，介绍一个新增动作序列的入库流程。

F1、服务器对该动作库以外的任一新增动作序列，获取该新增动作序列关联的新增参考音频和新增参考文本。

其中，该新增参考文本指示该新增参考音频的语义信息，该新增动作序列用于控制对应的样本形象执行配合该新增参考音频的动作。

步骤 F1 与步骤 501 同理，此处不再赘述。

F2、服务器基于该新增参考文本中词语和该新增参考音频中音素的关联关系，将该新增动作序列划分为多个新增动作片段。

其中，每个新增动作片段与该新增参考文本中的一个词语以及该新增参考音频中的一个音素相关联。

步骤 F2 与步骤 502 同理，此处不再赘述。

F3、服务器对每个新增动作片段，基于该新增动作片段的动作特征，从该动作库的多个动作集合中，确定该新增动作片段所属的目标动作集合。

在一些实施例中，对每个新增动作片段，基于步骤 503 同理的方式，计算该新增动作片段的动作特征，再计算该新增动作片段的动作特征与每个动作集合的类别特征之间的距离，将该新增动作片段分配给距离最近的目标动作集合。

F4、服务器将该新增动作片段添加至该目标动作集合，更新该类别特征和该贡献度分数，从该目标动作集合中，剔除贡献度分数符合该剔除条件的样本动作片段。

在一些实施例中，将该新增动作片段分配给该目标动作集合以后，由于目标动作集合中的样本数量发生变化，其聚类中心即类别特征必然需要重新计算，因此基于步骤 E1 同理的方式重新计算该类别特征，相应地，由于类别特征发生变化，每个样本动作片段（含新增动作片段）的贡献度分数也必然需要重新计算，因此基于步骤 E2 同理的方式重新计算该贡献度分数，再基于步骤 E3 同理的方式，按照计算得到的新的贡献度分数，继续剔除掉贡献度分数符合剔除条件的样本动作片段。

在一个示例中，如图 8 所示，图 8 是本申请实施例提供的一种新增动作片段的数据补充原理图，针对新增动作片段落入的目标动作集合，假设加入了 2 个新增动作片段，基于步骤 E2 同理的方式计算出来这 2 个新增动作片段的贡献度分数分别是 0.7 和 0.04，那么将这 2 个新增动作片段包含在内，对整个目标动作集合中的每个样本动作片段按照贡献度分数进行重排序（倒序），并剔除掉重排序以后位于末位的样本动作片段，例如剔除掉贡献度分数为 0.04 的末位的新增动作片段。

上述所有可选技术方案，能够采用任意结合形成本公开的可选实施例，在此不再一一赘述。

本申请实施例提供的方法，通过对样本动作序列，按照参考文本和参考音频的指导，划分成一系列样本动作片段，再使用聚类方式将样本动作片段划分到多个动作类别，每个动作类别具有一个动作集合来存放聚类到本动作类别的动作数据，这样能够构建一个完备多种动作类别的动作库，使得属于不同动作类别的动作数据在语义层面上区分开来，便于后续投入到动作生成流程中，以语义标签为索引来检测最匹配的动作类别，从而能够提升动作生成效率和准确率。

在以上的动作库建立方案中，提供了自动化的学习语义生产、自动分类和自动筛选的机制，方便了自动化剔除掉低质量的样本，并随时可以补充新的样本到任意动作类别，只需要利用贡献度分数对动作类别中的动作数据进行重新清洗，保证了动作库的高质量，也提升了

每个动作类别中动作数据的统一度。

图 9 是本申请实施例提供的一种虚拟形象的动作生成装置的结构示意图，如图 9 所示，该装置包括：

获取模块 901，用于获取虚拟形象的音频和文本，该文本指示该音频的语义信息；

分析模块 902，用于基于该文本，确定该文本的语义标签，该语义标签表征该文本中词语的词性信息或者该文本表达的情感信息中的至少一项；

检索模块 903，用于从预设动作库中，检索与该语义标签相匹配的动作类别和属于该动作类别的动作数据，该预设动作库包括该虚拟形象的、属于多种动作类别的动作数据；

生成模块 904，用于基于该动作数据，生成该虚拟形象的动作序列，该动作序列用于控制该虚拟形象执行配合该音频的动作。

本申请实施例提供的装置，通过以音频和文本作为双模态的驱动信号，在文本的基础上提取语义层面的语义标签，方便在预设动作库中检索到与语义标签匹配的动作类别，这个动作类别能够与音频的语义信息高度适配，反映出来虚拟形象在播报音频的情感倾向和潜在语义，进而检索属于该动作类别的动作数据，基于动作数据，为虚拟形象快速、高效地合成准确率更高的动作序列，不但提升虚拟形象的动作生成效率，且提升动作生成准确率。

进一步的，动作序列能够控制虚拟形象做出与音频在语义层面上配合的肢体动作，并非是简单跟随音频节奏进行律动，使得声画适配度、准确度极大提升，不会产生机械死板的视觉效果，能够提升虚拟形象的仿真度、拟人度，优化虚拟形象的渲染效果。

在一些实施例中，该分析模块 902 用于：基于该文本，确定该文本的情感标签；基于该文本，确定该文本中包含的至少一个词语；从词性表中查询每个该词语所属的词性标签；将该情感标签和该至少一个词语所属的词性标签，确定为该文本的语义标签。

在一些实施例中，该检索模块用于对该文本中包含的每个词语：基于该词语所属的语义标签，从该预设动作库中检索与该语义标签相匹配的动作类别；从该预设动作库中检索属于该动作类别的动作数据。

在一些实施例中，基于图 9 的装置组成，该生成模块 904 包括：

确定单元，用于对该文本中包含的每个词语：基于该词语关联的音素，从该音频中确定该音素所属的音频片段；

片段生成单元，用于基于该词语对应的该动作数据和该音频片段，生成与该音频片段相匹配的动作片段；

序列生成单元，用于基于每个词语的音频片段相匹配的每个动作片段，生成与该音频相匹配的该动作序列。

在一些实施例中，基于图 9 的装置组成，该片段生成单元包括：

确定子单元，用于从该动作数据中，确定与该词语的语义匹配度最高的该至少一个关键动作帧；

合成子单元，用于基于该音频片段，将该至少一个关键动作帧合成为与该音频片段相匹配的该动作片段。

在一些实施例中，该合成子单元用于：在该关键动作帧的帧数不超过该音频片段的音频帧数的情况下，对该至少一个关键动作帧进行插帧，得到与该音频片段等长的该动作片段；在该关键动作帧的帧数超过该音频帧数的情况下，创建与该音频片段等长的动作片段，将该动作片段的每一帧填充为预设动作类别下的预设动作帧。

在一些实施例中，该序列生成单元用于：基于每个音频片段的时间戳顺序，拼接每个音频片段相匹配的每个动作片段，得到拼接动作序列；对该拼接动作序列中的每个动作帧进行动作平滑，得到该动作序列。

在一些实施例中，该检索模块 903 用于：提取每个该语义标签的语义特征；查询该预设动作库中多个候选类别的类别特征；从该多个候选类别中确定该动作类别，该动作类别的该

类别特征与该语义特征符合相似条件。

在一些实施例中，该检索模块 903 还用于：在该多个候选类别的类别特征与该语义特征均不符合该相似条件的情况下，将与该语义标签相匹配的动作类别配置为预设动作类别。

上述所有可选技术方案，能够采用任意结合形成本公开的可选实施例，在此不再一一赘述。

需要说明的是：上述实施例提供的虚拟形象的动作生成装置在生成虚拟形象的肢体动作时，仅以上述各功能模块的划分进行举例说明，实际应用中，能够根据需要而将上述功能分配由不同的功能模块完成，即将计算机设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的虚拟形象的动作生成装置与虚拟形象的动作生成方法实施例属于同一构思，其具体实现过程详见虚拟形象的动作生成方法实施例，这里不再赘述。

图 10 是本申请实施例提供的一种虚拟形象的动作库的构建装置的结构示意图，如图 10 所示，该装置包括：

样本获取模块 1001，用于获取每个样本形象的样本动作序列、参考音频和参考文本，该参考文本指示该参考音频的语义信息，该样本动作序列用于控制该样本形象执行配合该参考音频的动作；

片段划分模块 1002，用于基于该参考文本中词语和该参考音频中音素的关联关系，将该样本动作序列划分为多个样本动作片段，每个样本动作片段与该参考文本中的一个词语以及该参考音频中的一个音素相关联；

聚类模块 1003，用于基于该样本动作片段的动作特征，对每个样本形象的每个样本动作片段进行聚类，得到多个动作集合，每个动作集合指示聚类到同一动作类别下不同样本形象的动作数据；

构建模块 1004，用于基于该多个动作集合，构建动作库。

本申请实施例提供的装置，通过对样本动作序列，按照参考文本和参考音频的指导，划分成一系列样本动作片段，再使用聚类方式将样本动作片段划分到多个动作类别，每个动作类别具有一个动作集合来存放聚类到本动作类别的全部动作数据，这样能够构建一个完备多种动作类别下的动作库，使得属于不同动作类别的动作数据在语义层面上区分开来，便于后续投入到动作生成流程中，以语义标签为索引来检测最匹配的动作类别，从而能够提升动作生成效率和准确率。

在一些实施例中，该片段划分模块 1002 用于：对该参考文本中的每个词语，基于该词语关联的音素，从该样本音频中确定该音素所关联的该样本音频片段；基于每个样本音频片段的时间戳区间，将该样本动作序列划分成多个样本动作片段，每个样本动作片段与一个样本音频片段的时间戳区间对齐。

在一些实施例中，基于图 10 的装置组成，该装置还包括：

特征获取模块，用于对每个动作集合，获取该动作集合所指示的动作类别的类别特征，该类别特征为该动作集合中每个样本动作片段的平均动作特征；

确定模块，用于确定该动作集合中每个样本动作片段的动作特征对该类别特征的贡献度分数，该贡献度分数表征该样本动作片段与该动作类别的匹配程度；

剔除模块，用于从该动作集合中，剔除贡献度分数符合剔除条件的样本动作片段；

迭代模块，用于基于剔除后的动作集合，更新该类别特征和该贡献度分数，迭代多次执行剔除操作，在满足迭代停止条件的情况下停止迭代。

在一些实施例中，该确定模块用于：对该动作集合中任一样本动作片段，获取除了该样本动作片段以外的每个其余动作片段的动作分数，该动作分数表征该其余动作片段与该类别特征的相似程度；基于每个其余动作片段的动作分数，确定排除该样本动作片段以后的类内方差，将该类内方差确定为该样本动作片段的贡献度分数。

在一些实施例中，该剔除模块用于：按照贡献度分数从大到小的顺序，对该动作集合中的每个样本动作片段进行排序，剔除在该排序中位于末位的样本动作片段。

在一些实施例中，该样本获取模块 1001 还用于：对该动作库以外的任一新增动作序列，获取该新增动作序列关联的新增参考音频和新增参考文本；

该片段划分模块 1002 还用于：基于该新增参考文本中词语和该新增参考音频中音素的关联关系，将该新增动作序列划分为多个新增动作片段；

该聚类模块 1003 还用于：对每个新增动作片段，基于该新增动作片段的动作特征，从该动作库的多个动作集合中，确定该新增动作片段所属的目标动作集合；

该构建模块 1004 还用于：将该新增动作片段添加至该目标动作集合，更新该类别特征和该贡献度分数，从该目标动作集合中，剔除贡献度分数符合该剔除条件的样本动作片段。

上述所有可选技术方案，能够采用任意结合形成本公开的可选实施例，在此不再一一赘述。

需要说明的是：上述实施例提供的虚拟形象的动作库的构建装置在构建动作库时，仅以上述各功能模块的划分进行举例说明，实际应用中，能够根据需要而将上述功能分配由不同的功能模块完成，即将计算机设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的虚拟形象的动作库的构建装置与虚拟形象的动作库的构建方法实施例属于同一构思，其具体实现过程详见虚拟形象的动作库的构建方法实施例，这里不再赘述。

图 11 是本申请实施例提供的一种计算机设备的结构示意图，如图 11 所示，该计算机设备 1100 可因配置或性能不同而产生比较大的差异，该计算机设备 1100 包括一个或一个以上处理器（Central Processing Units，CPU）1101 和一个或一个以上的存储器 1102，其中，该存储器 1102 中存储有至少一条计算机程序，该至少一条计算机程序由该一个或一个以上处理器 1101 加载并执行以实现上述各个实施例提供的虚拟形象的动作生成方法或虚拟形象的动作库的构建方法。可选地，该计算机设备 1100 还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，该计算机设备 1100 还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，例如包括至少一条计算机程序的存储器，上述至少一条计算机程序可由计算机设备中的处理器执行以完成上述各个实施例中的虚拟形象的动作生成方法或虚拟形象的动作库的构建方法。例如，该计算机可读存储介质包括 ROM（Read-Only Memory，只读存储器）、RAM（Random-Access Memory，随机存取存储器）、CD-ROM（Compact Disc Read-Only Memory，只读光盘）、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产品，包括一条或多条计算机程序，该一条或多条计算机程序存储在计算机可读存储介质中。计算机设备的一个或多个处理器能够从计算机可读存储介质中读取该一条或多条计算机程序，该一个或多个处理器执行该一条或多条计算机程序，使得计算机设备能够执行以完成上述实施例中的虚拟形象的动作生成方法或虚拟形象的动作库的构建方法。

本领域普通技术人员能够理解实现上述实施例的全部或部分步骤能够通过硬件来完成，也能够通过程序来指令相关的硬件完成，可选地，该程序存储于一种计算机可读存储介质中，可选地，上述提到的存储介质是只读存储器、磁盘或光盘等。

以上所述仅为本申请的可选实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权 利 要 求 书

1.一种虚拟形象的动作生成方法，应用于计算机设备，所述方法包括：

获取虚拟形象的音频和文本，所述文本指示所述音频的语义信息；

基于所述文本，确定所述文本的语义标签，所述语义标签表征所述文本中词语的词性信息或者所述文本表达的情感信息中的至少一项；

从预设动作库中，检索与所述语义标签相匹配的动作类别和属于所述动作类别的动作数据，所述预设动作库包括所述虚拟形象的、属于多种动作类别的动作数据；

基于所述动作数据，生成所述虚拟形象的动作序列，所述动作序列用于控制所述虚拟形象执行配合所述音频的动作。

2.根据权利要求 1 所述的方法，其中，所述基于所述文本，确定所述文本的语义标签包括：

基于所述文本，确定所述文本的情感标签；

基于所述文本，确定所述文本中包含的至少一个词语；

从词性表中查询每个所述词语所属的词性标签；

将所述情感标签和所述至少一个词语所属的词性标签，确定为所述文本的语义标签。

3.根据权利要求 1 所述的方法，其中，所述从预设动作库中，检索与所述语义标签相匹配的动作类别和属于所述动作类别的动作数据，包括：

对所述文本中包含的每个词语：

基于所述词语所属的语义标签，从所述预设动作库中检索与所述语义标签相匹配的动作类别；

从所述预设动作库中检索属于所述动作类别的动作数据。

4.根据权利要求 3 所述的方法，其中，所述基于所述动作数据，生成所述虚拟形象的动作序列包括：

对所述文本中包含的每个词语：基于所述词语关联的音素，从所述音频中确定所述音素所属的音频片段，以及，基于所述词语对应的所述动作数据和所述音频片段，生成与所述音频片段相匹配的动作片段；

基于每个词语的音频片段相匹配的每个动作片段，生成与所述音频相匹配的所述动作序列。

5.根据权利要求 4 所述的方法，其中，所述基于所述词语对应的所述动作数据和所述音频片段，生成与所述音频片段相匹配的动作片段包括：

从所述动作数据中，确定与所述词语的语义匹配度最高的至少一个关键动作帧；

基于所述音频片段，将所述至少一个关键动作帧合成为与所述音频片段相匹配的所述动作片段。

6.根据权利要求 5 所述的方法，其中，所述基于所述音频片段，将所述至少一个关键动作帧合成为与所述音频片段相匹配的所述动作片段包括：

在所述关键动作帧的帧数不超过所述音频片段的音频帧数的情况下，对所述至少一个关键动作帧进行插帧，得到与所述音频片段等长的所述动作片段；

在所述关键动作帧的帧数超过所述音频帧数的情况下，创建与所述音频片段等长的动作片段，将所述动作片段的每一帧填充为预设动作类别下的预设动作帧。

7.根据权利要求 4 所述的方法，其中，所述每个词语的音频片段相匹配的每个动作片段，生成与所述音频相匹配的所述动作序列包括：

基于每个音频片段的时间戳顺序，拼接每个音频片段相匹配的每个动作片段，得到拼接动作序列；

对所述拼接动作序列中的每个动作帧进行动作平滑，得到所述动作序列。

8.根据权利要求 1 所述的方法，其中，所述从预设动作库中，检索与所述语义标签相匹

配的动作类别和属于所述动作类别的动作数据包括：

提取所述语义标签的语义特征；

查询所述预设动作库中多个候选类别的类别特征；

从所述多个候选类别中确定所述动作类别，所述动作类别的所述类别特征与所述语义特征符合相似条件。

9.根据权利要求 8 所述的方法，其中，所述方法还包括：

在所述多个候选类别的类别特征与所述语义特征均不符合所述相似条件的情况下，将与所述语义标签相匹配的动作类别配置为预设动作类别。

10.一种虚拟形象的动作库的构建方法，应用于计算机设备，所述方法包括：

获取每个样本形象的样本动作序列、参考音频和参考文本，所述参考文本指示所述参考音频的语义信息，所述样本动作序列用于控制所述样本形象执行配合所述参考音频的动作；

基于所述参考文本中词语和所述参考音频中音素的关联关系，将所述样本动作序列划分为多个样本动作片段，每个样本动作片段与所述参考文本中的一个词语以及所述参考音频中的一个音素相关联；

基于所述样本动作片段的动作特征，对每个样本形象的每个样本动作片段进行聚类，得到多个动作集合，每个动作集合指示属于同一动作类别且属于不同样本形象的动作数据；

基于所述多个动作集合，构建动作库。

11.根据权利要求 10 所述的方法，其中，所述基于所述参考文本中词语和所述参考音频中音素的关联关系，将所述样本动作序列划分为多个样本动作片段包括：

对所述参考文本中的每个词语，基于所述词语关联的音素，从所述样本音频中确定所述音素所关联的所述样本音频片段；

基于每个样本音频片段的时间戳区间，将所述样本动作序列划分成多个样本动作片段，每个样本动作片段与一个样本音频片段的时间戳区间对齐。

12.根据权利要求 10 所述的方法，其中，所述方法还包括：

对每个动作集合，获取所述动作集合所指示的动作类别的类别特征，所述类别特征为所述动作集合中每个样本动作片段的平均动作特征；

确定所述动作集合中每个样本动作片段的动作特征对所述类别特征的贡献度分数，所述贡献度分数表征所述样本动作片段与所述动作类别的匹配程度；

从所述动作集合中，剔除贡献度分数符合剔除条件的样本动作片段；

基于剔除后的动作集合，更新所述类别特征和所述贡献度分数，迭代多次执行剔除操作，在满足迭代停止条件的情况下停止迭代。

13.根据权利要求 12 所述的方法，其中，所述确定所述动作集合中每个样本动作片段的动作特征对所述类别特征的贡献度分数包括：

对所述动作集合中任一样本动作片段，获取除了所述样本动作片段以外的每个其余动作片段的动作分数，所述动作分数表征所述其余动作片段与所述类别特征的相似程度；

基于每个其余动作片段的动作分数，确定排除所述样本动作片段以后的类内方差，将所述类内方差确定为所述样本动作片段的贡献度分数。

14.根据权利要求 12 所述的方法，其中，所述从所述动作集合中，剔除贡献度分数符合剔除条件的样本动作片段，得到动作集合包括：

按照贡献度分数从大到小的顺序，对所述动作集合中的每个样本动作片段进行排序，剔除在所述排序中位于末位的样本动作片段。

15.根据权利要求 12 所述的方法，其中，所述方法还包括：

对所述预设动作库以外的任一新增动作序列，获取所述新增动作序列关联的新增参考音频和新增参考文本；

基于所述新增参考文本中词语和所述新增参考音频中音素的关联关系，将所述新增动作

序列划分为多个新增动作片段；

对每个新增动作片段，基于所述新增动作片段的动作特征，从所述预设动作库的多个动作集合中，确定所述新增动作片段所属的目标动作集合；

将所述新增动作片段添加至所述目标动作集合，更新所述类别特征和所述贡献度分数，从所述目标动作集合中，剔除贡献度分数符合所述剔除条件的样本动作片段。

16.一种虚拟形象的动作生成装置，所述装置包括：

获取模块，用于获取虚拟形象的音频和文本，所述文本指示所述音频的语义信息；

分析模块，用于基于所述文本，确定所述文本的语义标签，所述语义标签表征所述文本中词语的词性信息或者所述文本表达的情感信息中的至少一项；

检索模块，用于从预设动作库中，检索与所述语义标签相匹配的动作类别和属于所述动作类别的动作数据，所述预设动作库包括所述虚拟形象的、属于多种动作类别的动作数据；

生成模块，用于基于所述动作数据，生成所述虚拟形象的动作序列，所述动作序列用于控制所述虚拟形象执行配合所述音频的动作。

17.一种虚拟形象的动作库的构建装置，所述装置包括：

样本获取模块，用于获取每个样本形象的样本动作序列、参考音频和参考文本，所述参考文本指示所述参考音频的语义信息，所述样本动作序列用于控制所述样本形象执行配合所述参考音频的动作；

片段划分模块，用于基于所述参考文本中词语和所述参考音频中音素的关联关系，将所述样本动作序列划分为多个样本动作片段，每个样本动作片段与所述参考文本中的一个词语以及所述参考音频中的一个音素相关联；

聚类模块，用于基于所述样本动作片段的动作特征，对每个样本形象的每个样本动作片段进行聚类，得到多个动作集合，每个动作集合指示属于同一动作类别且属于不同样本形象的动作数据；

构建模块，用于基于所述多个动作集合，构建动作库。

18.一种计算机设备，所述计算机设备包括一个或多个处理器和一个或多个存储器，所述一个或多个存储器中存储有至少一条计算机程序，所述至少一条计算机程序由所述一个或多个处理器加载并执行以实现如权利要求 1 至权利要求 9 任一项所述的虚拟形象的动作生成方法；或，如权利要求 10 至权利要求 15 任一项所述的虚拟形象的动作库的构建方法。

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计算机程序，所述至少一条计算机程序由处理器加载并执行以实现如权利要求 1 至权利要求 9 任一项所述的虚拟形象的动作生成方法；或，如权利要求 10 至权利要求 15 任一项所述的虚拟形象的动作库的构建方法。

20.一种计算机程序产品，所述计算机程序产品包括至少一条计算机程序，所述至少一条计算机程序由处理器加载并执行以实现如权利要求 1 至权利要求 9 任一项所述的虚拟形象的动作生成方法；或，如权利要求 10 至权利要求 15 任一项所述的虚拟形象的动作库的构建方法。


102
101
有线网络或
无线网络

图 1

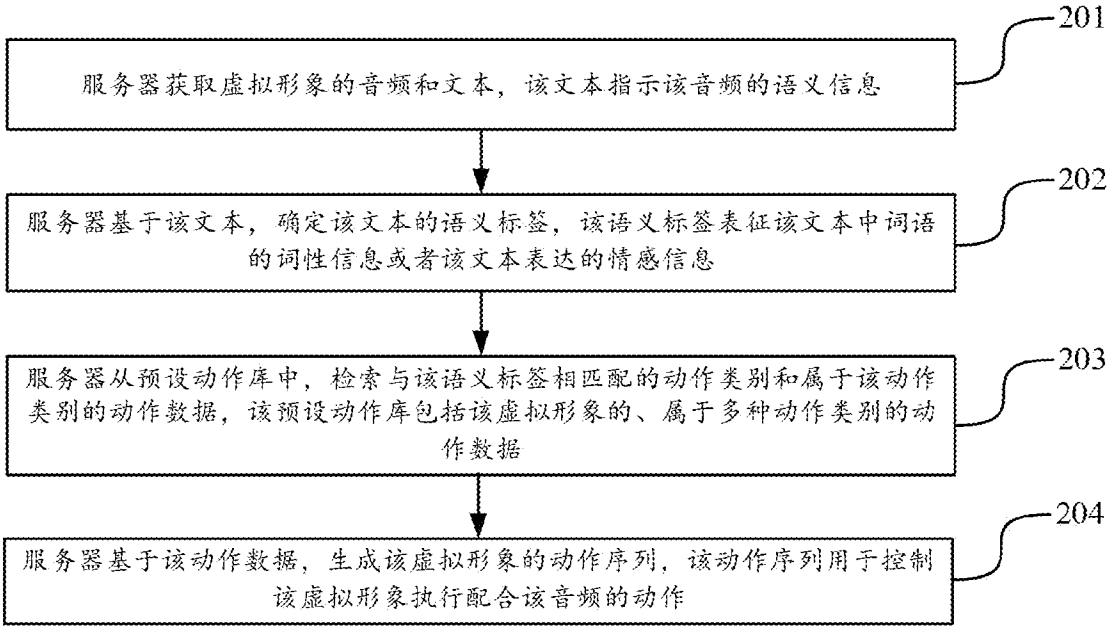
201
服务器获取虚拟形象的音频和文本，该文本指示该音频的语义信息
202
服务器基于该文本，确定该文本的语义标签，该语义标签表征该文本中词语的词性信息或者该文本表达的情感信息
203
服务器从预设动作库中，检索与该语义标签相匹配的动作类别和属于该动作类别的动作数据，该预设动作库包括该虚拟形象的、属于多种动作类别的动作数据
204
服务器基于该动作数据，生成该虚拟形象的动作序列，该动作序列用于控制该虚拟形象执行配合该音频的动作

图 2

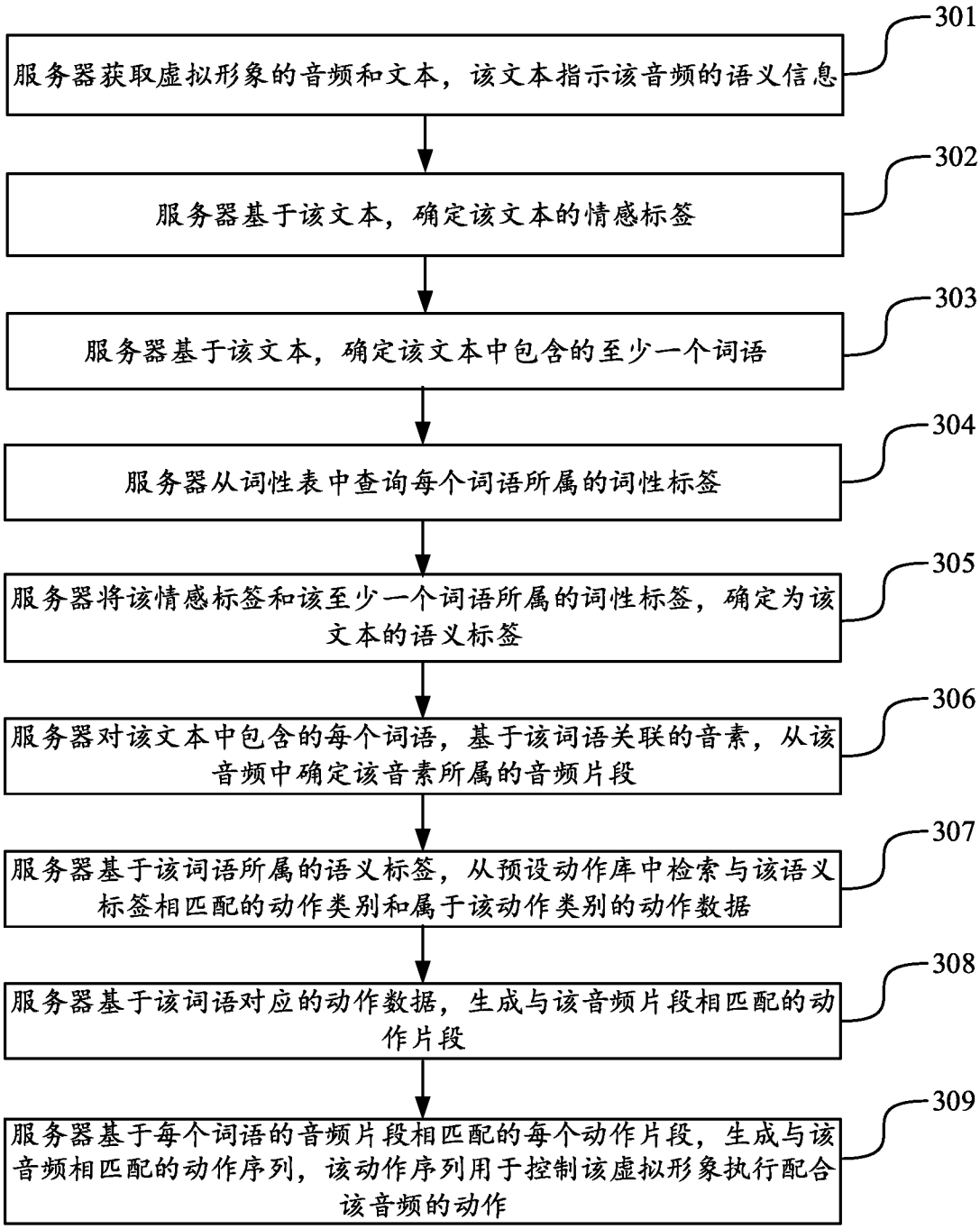
服务器获取虚拟形象的音频和文本，该文本指示该音频的语义信息
301
服务器基于该文本，确定该文本的情感标签
302
服务器基于该文本，确定该文本中包含的至少一个词语
303
服务器从词性表中查询每个词语所属的词性标签
304
服务器将该情感标签和该至少一个词语所属的词性标签，确定为该文本的语义标签
305
服务器对该文本中包含的每个词语，基于该词语关联的音素，从该音频中确定该音素所属的音频片段
306
服务器基于该词语所属的语义标签，从预设动作库中检索与该语义标签相匹配的动作类别和属于该动作类别的动作数据
307
服务器基于该词语对应的动作数据，生成与该音频片段相匹配的动作片段
308
服务器基于每个词语的音频片段相匹配的每个动作片段，生成与该音频相匹配的动作序列，该动作序列用于控制该虚拟形象执行配合该音频的动作
309

图 3

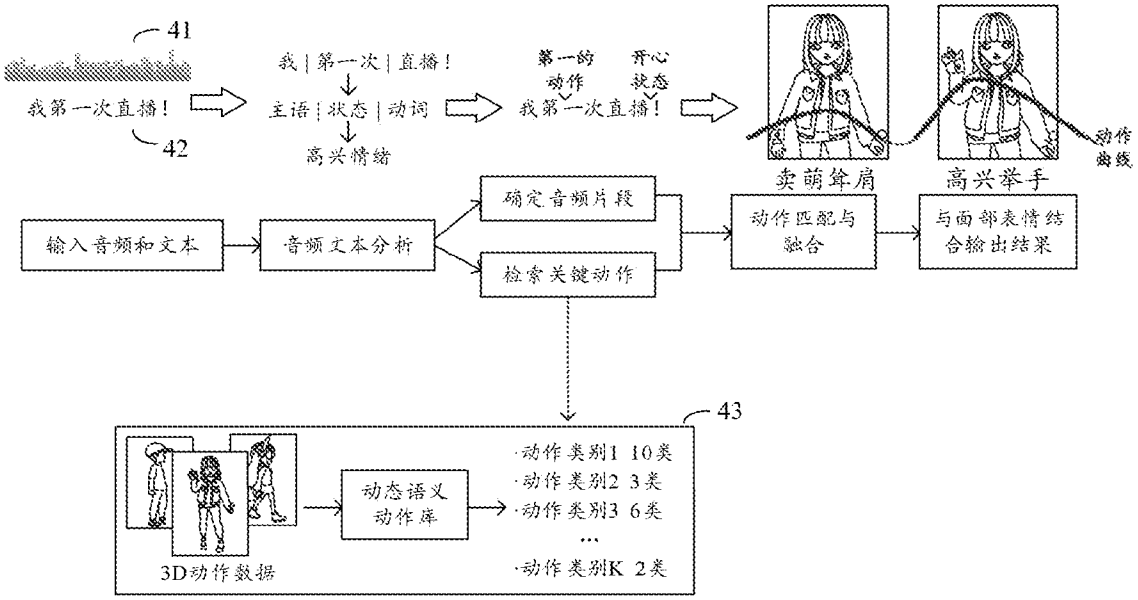
41
我第一次直播！
42
我 | 第一次 | 直播！
主语 | 状态 | 动词
高兴情绪
第一的 动作
开心 状态
我第一次直播！
卖萌耸肩
高兴举手
动作曲线
输入音频和文本
音频文本分析
确定音频片段
检索关键动作
动作匹配与融合
与面部表情结合输出结果
43
3D动作数据
动态语义动作库
·动作类别1 10类
·动作类别2 3类
·动作类别3 6类
…
·动作类别K 2类

图 4

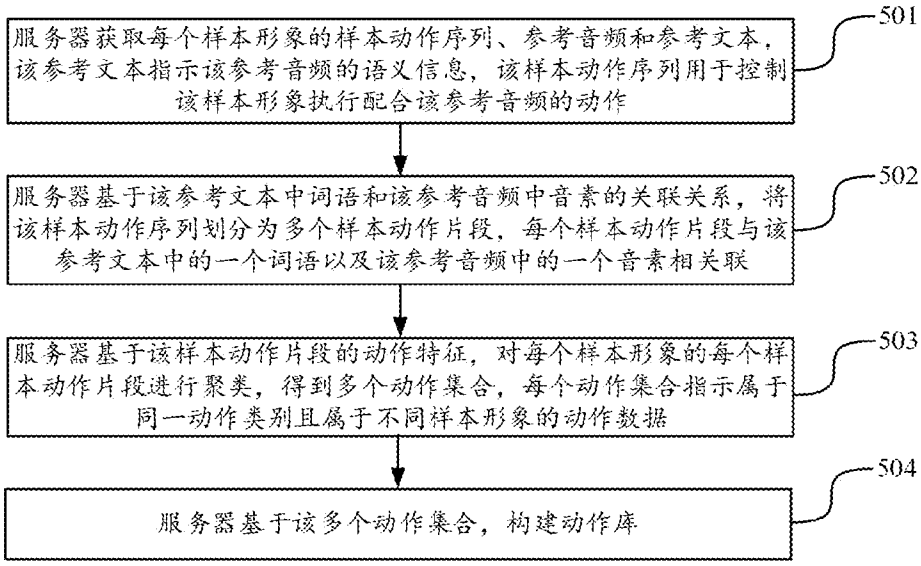
服务器获取每个样本形象的样本动作序列、参考音频和参考文本，该参考文本指示该参考音频的语义信息，该样本动作序列用于控制该样本形象执行配合该参考音频的动作
501
服务器基于该参考文本中词语和该参考音频中音素的关联关系，将该样本动作序列划分为多个样本动作片段，每个样本动作片段与该参考文本中的一个词语以及该参考音频中的一个音素相关联
502
服务器基于该样本动作片段的动作特征，对每个样本形象的每个样本动作片段进行聚类，得到多个动作集合，每个动作集合指示属于同一动作类别且属于不同样本形象的动作数据
503
服务器基于该多个动作集合，构建动作库
504

图 5

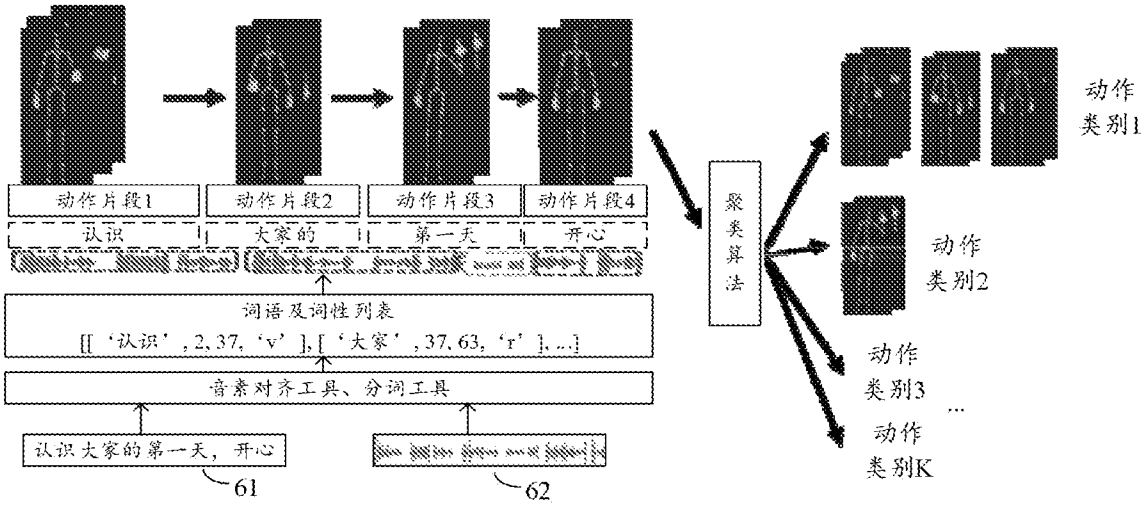
动作片段1
动作片段2
动作片段3
动作片段4
认识
大家的
第一天
开心
词语及词性列表
[['认识' , 2, 37, 'v'], ['大家' , 37, 63, 'r'], ...]
音素对齐工具、分词工具
认识大家的第一天，开心
61
62
聚类算法
动作类别1
动作类别2
动作类别3
...
动作类别K

图 6

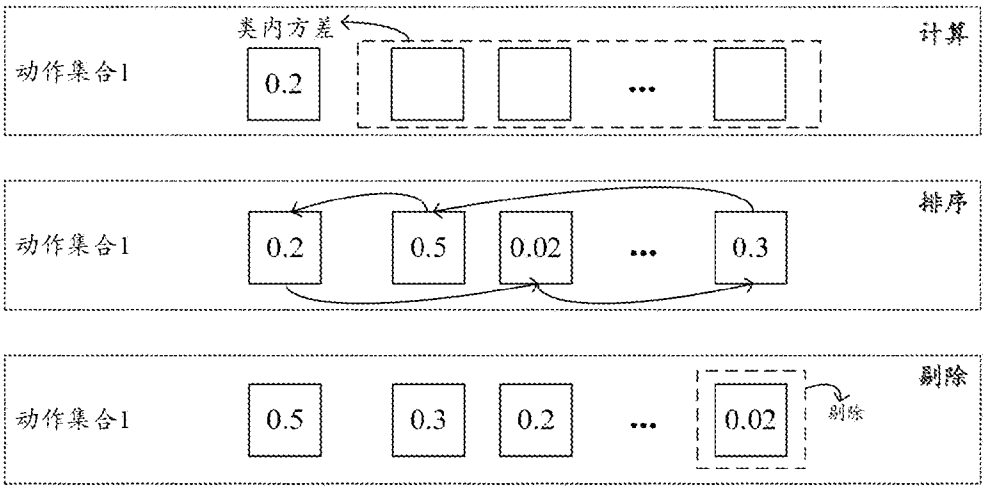
动作集合1
类内方差
0.2
计算
排序
0.5
0.02
0.3
剔除
删除

图 7

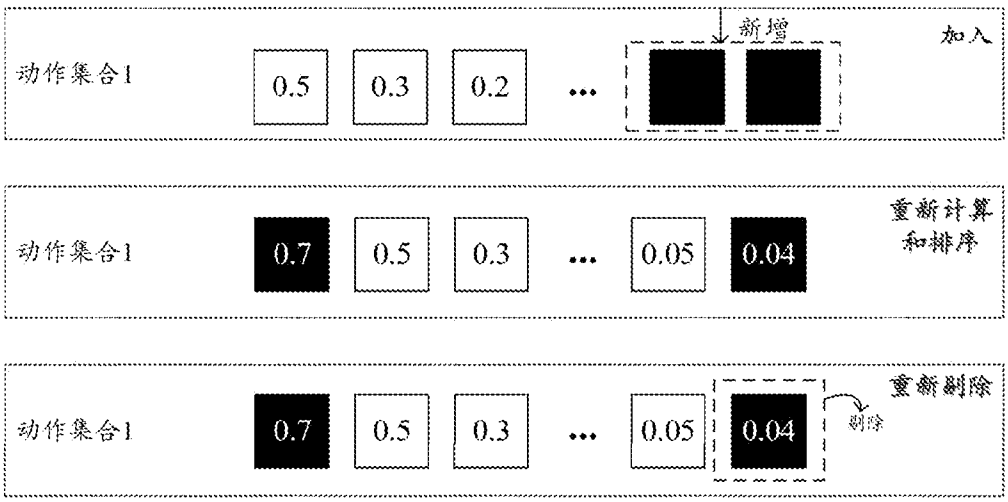
动作集合1
新增
加入
0.5
0.3
0.2
重新计算和排序
0.7
0.05
0.04
重新剔除
删除

图 8

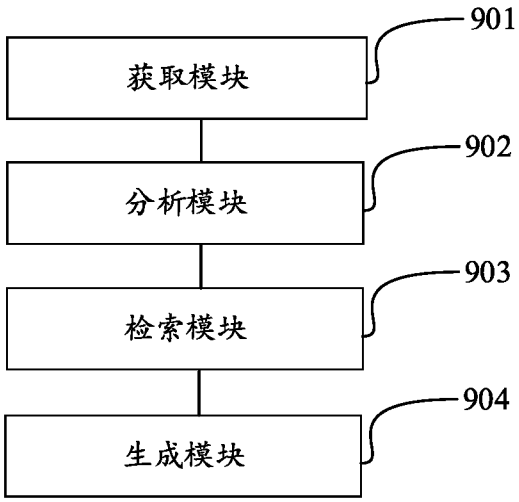
901
获取模块
902
分析模块
903
检索模块
904
生成模块

图 9

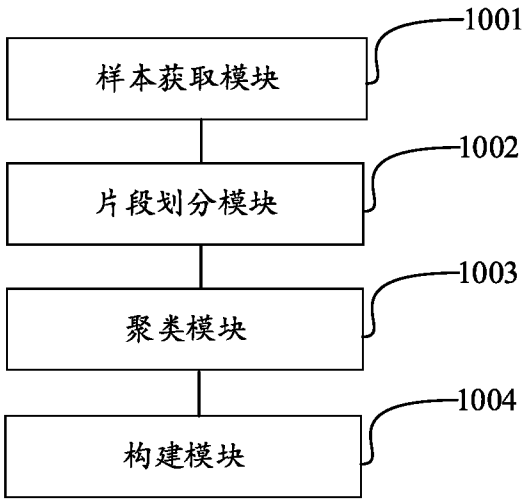
1001
样本获取模块
1002
片段划分模块
1003
聚类模块
1004
构建模块

图 10

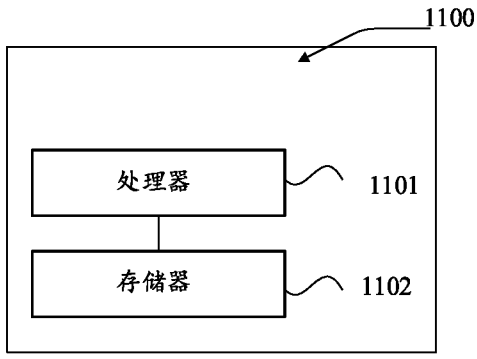
1100
处理器
1101
存储器
1102

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2024/093505

A. CLASSIFICATION OF SUBJECT MATTER

G06T13/40(2011.01)i; G06T13/20(2011.01)i; G06F16/907(2019.01)i; G06F16/906(2019.01)i; G06F40/30(2020.01)i; G06F40/289(2020.01)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:G06T,G06F

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: 虚拟, 动画, 动作, 表情, 情绪, 音频, 语音, 文本, 文字, 库, 音素, virtual, animation, action, expression, mood, audio, speech, text, library, phoneme

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 116958342 A (TENCENT TECHNOLOGY (SHENZHEN) CO., LTD.) 27 October 2023 (2023-10-27) description, paragraphs 0005-0069	1-20
X	CN 115147521 A (BEIJING ZHONGKE SHIWEI CULTURE TECHNOLOGY CO., LTD.) 04 October 2022 (2022-10-04) description, paragraphs 0025-0052 and 0079-0085	1-3, 8-9, 16, 18-20
X	CN 114911973 A (NETEASE (HANGZHOU) NETWORK CO., LTD.) 16 August 2022 (2022-08-16) description, paragraphs 0102-0108	10-11, 17-20
A	CN 112650831 A (BEIJING DAMI TECHNOLOGY CO., LTD.) 13 April 2021 (2021-04-13) entire document	1-20
A	CN 114513678 A (ALIBABA GROUP HOLDING LTD.) 17 May 2022 (2022-05-17) entire document	1-20
A	US 2004064321 A1 (AT & T CORP.) 01 April 2004 (2004-04-01) 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
24 July 2024	**29 July 2024**

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/CN2024/093505

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 116958342 A	27 October 2023	None	
CN 115147521 A	04 October 2022	None	
CN 114911973 A	16 August 2022	None	
CN 112650831 A	13 April 2021	None	
CN 114513678 A	17 May 2022	None	
US 2004064321 A1	01 April 2004	None	

国际检索报告

国际申请号
PCT/CN2024/093505

A. 主题的分类

G06T13/40(2011.01)i；G06T13/20(2011.01)i；G06F16/907(2019.01)i；G06F16/906(2019.01)i；G06F40/30(2020.01)i；G06F40/289(2020.01)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:G06T,G06F

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：虚拟，动画，动作，表情，情绪，音频，语音，文本，文字，库，音素，virtual, animation, action, expression, mood, audio, speech, text, library, phoneme

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 116958342 A (腾讯科技（深圳）有限公司) 2023年10月27日 (2023 - 10 - 27) 说明书第0005-0069段	1-20
X	CN 115147521 A (北京中科视维文化科技有限公司) 2022年10月4日 (2022 - 10 - 04) 说明书第0025-0052、0079-0085段	1-3,8-9,16,18-20
X	CN 114911973 A (网易（杭州）网络有限公司) 2022年8月16日 (2022 - 08 - 16) 说明书第0102-0108段	10-11,17-20
A	CN 112650831 A (北京大米科技有限公司) 2021年4月13日 (2021 - 04 - 13) 全文	1-20
A	CN 114513678 A (阿里巴巴集团控股有限公司) 2022年5月17日 (2022 - 05 - 17) 全文	1-20
A	US 2004064321 A1 (AT & T CORP.) 2004年4月1日 (2004 - 04 - 01) 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
2024年7月24日	2024年7月29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 邱爽 电话号码 (+86) 010-53961362

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2024/093505

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	116958342	A	2023年10月27日	无	
CN	115147521	A	2022年10月4日	无	
CN	114911973	A	2022年8月16日	无	
CN	112650831	A	2021年4月13日	无	
CN	114513678	A	2022年5月17日	无	
US	2004064321	A1	2004年4月1日	无	