(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
2021 年 3 月 11 日 (11.03.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/043053 A1

(51) 国际专利分类号:
G06F 3/01 (2006.01)

(21) 国际申请号: PCT/CN2020/111615

(22) 国际申请日: 2020 年 8 月 27 日 (27.08.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910824770.0 2019年9月2日 (02.09.2019) 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: **暴林超 (BAO, Linchao)**; 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)。 **康世胤 (KANG, Shiyin)**; 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)。 **王盛 (WANG, Sheng)**; 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)。 **林祥凯 (LIN, Xiangkai)**; 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)。 **季兴 (JI, Xing)**; 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)。 **朱展图 (ZHU, Zhantu)**; 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)。 **李广之 (LEI, Kuongchi)**; 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)。 **陀得意 (TUO, Deyi)**; 中国广东省

(54) **Title:** ANIMATION IMAGE DRIVING METHOD BASED ON ARTIFICIAL INTELLIGENCE, AND RELATED DEVICE

(54) **发明名称:** 一种基于人工智能的动画形象驱动方法和相关装置

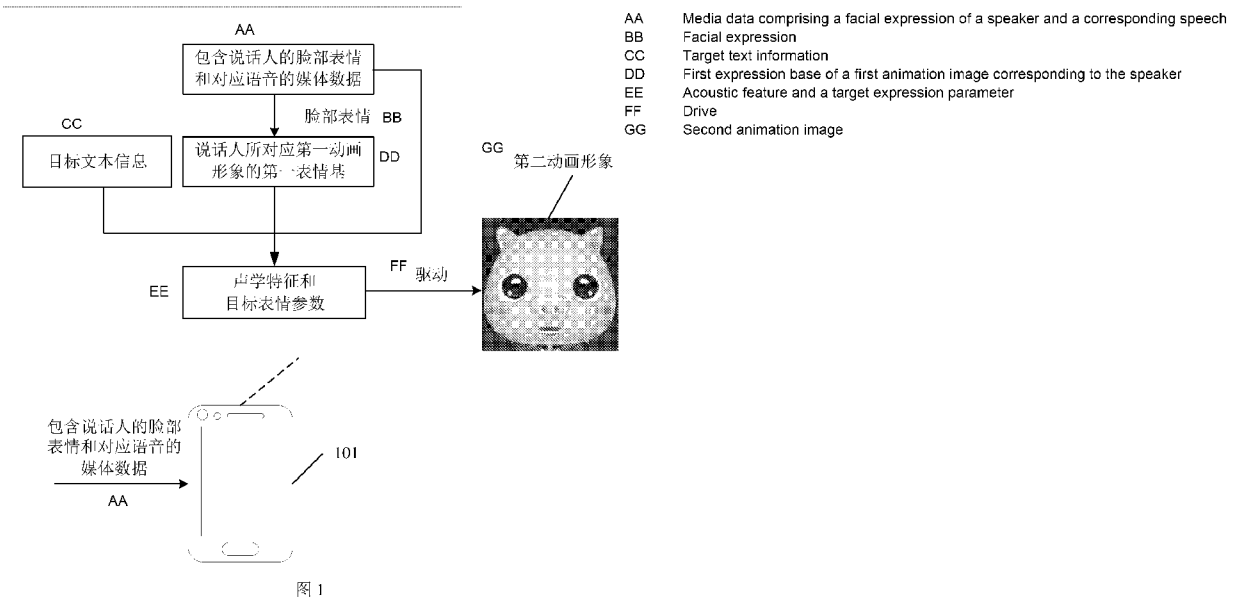


图 1

(57) **Abstract:** An animation image driving method based on artificial intelligence, and a related device. The method comprises: collecting media data of facial expression changes when a speaker speaks a speech, and determining a first expression base of a first animation image corresponding to the speaker, wherein the first expression base can reflect different expressions of the first animation image; after target text information used for driving a second animation image is determined, determining an acoustic feature and a target expression parameter corresponding to the target text information according to the target text information, the collected media data and the first expression base; and driving the second animation image having a second expression base by means of the acoustic feature and the target expression parameter, so that the second animation image can give out the sound of the target text information

WO 2021/043053 A1

[见续页]

深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 **刘朋 (LIU, Peng)**; 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: 深圳市深佳知识产权代理事务所（普通合伙）(**SHENPAT INTELLECTUAL PROPERTY AGENCY**); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦B座18C2、18D、18E、18E2, Guangdong 518001 (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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, 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))。

spoken by the speaker by means of acoustic feature simulation, and a facial expression conforming to the due expression of the speaker is made in a sounding process, so that vivid substitution feeling and immersion feeling are brought to a user, and the interaction experience of the user and the animation image is improved.

(57) 摘要: 一种基于人工智能的动画形象驱动方法和相关装置，采集说话人说出语音时脸部表情变化的媒体数据，确定说话人所对应第一动画形象的第一表情基，第一表情基可以体现第一动画形象的不同表情。在确定出用于驱动第二动画形象的目标文本信息后，根据目标文本信息、前述采集的媒体数据和第一表情基，确定对应目标文本信息的声学特征和目标表情参数。通过声学特征和目标表情参数，可以驱动具有第二表情基的第二动画形象，使得第二动画形象可以通过声学特征模拟发出说话人说出目标文本信息的声音，并且在发声过程中做出符合该说话人应有表情的脸部表情，给用户带来逼真的代入感和沉浸感，提高了用户与动画形象进行交互的体验。

一种基于人工智能的动画形象驱动方法和相关装置

本申请要求于2019年9月2日提交中国专利局、申请号201910824770.0、申请名称为“一种基于人工智能的动画形象驱动方法和装置”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

本申请涉及数据处理领域，特别是涉及动画形象驱动。

背景技术

随着计算机技术的发展，人机交互已经比较常见，但多为单纯的语音交互，例如，交互设备可以根据用户输入的文字或语音确定回复内容，并播放根据回复内容合成的虚拟声音。

这种类型的人机交互带来的用户沉浸感难以满足目前用户的交互需求，为了提高用户沉浸感，具有表情变化能力例如可以口型变化的动画形象作为与用户进行交互的交互对象属于目前的研发方向。

然而，目前并没有完善的动画形象驱动方式。

发明内容

为了解决上述技术问题，本申请提供了一种基于人工智能的动画形象驱动方法和装置，给用户带来逼真的代入感和沉浸感，提高了用户与动画形象进行交互的体验。

本申请实施例公开了如下技术方案：

第一方面，本申请实施例提供一种动画形象驱动方法，所述方法由音视频处理设备执行，所述方法包括：

获取包含说话人的脸部表情和对应语音的媒体数据；

根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；

根据目标文本信息、所述媒体数据和所述第一表情基，确定对应所述目标文本信息的声学特征和目标表情参数；所述声学特征用于标识模拟所述说话人说出所述目标文本信息的声音，所述目标表情参数用于标识模拟所述说话人说出所述目标文本信息的脸部表情相对于所述第一表情基的变化程度；

根据所述声学特征和所述目标表情参数，驱动具有第二表情基的第二动画形象。

第二方面，本申请实施例提供一种动画形象驱动装置，所述装置部署在音视频处理设备上，所述装置包括获取单元、第一确定单元、第二确定单元和驱动单元：

所述获取单元，用于获取包含说话人的脸部表情和对应语音的媒体数据；

所述第一确定单元，用于根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；

所述第二确定单元，用于根据目标文本信息、所述媒体数据和所述第一表情基，确定对应所述目标文本信息的声学特征和目标表情参数；所述声学特征用于标识模拟所述说话人说出所述目标文本信息的声音，所述目标表情参数用于标识模拟所述说话人说出所述目标文本信息的脸部表情相对于所述第一表情基的变化程度；

所述驱动单元，用于根据所述声学特征和所述目标表情参数，驱动具有第二表情基的第二动画形象。

第三方面，本申请实施例提供一种动画形象驱动方法，所述方法由音视频处理设备执行，所述方法包括：

获取包含说话人的脸部表情和对应语音的第一媒体数据；

根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；所述第一表情基的维数为第一维数，顶点拓扑为第一顶点拓扑；

根据所述第一表情基和待驱动的第二动画形象的第二表情基，确定目标表情基；所述第二表情基的维数为第二维数，顶点拓扑为第二顶点拓扑，所述目标表情基为具有第二顶点拓扑的第一动画形象对应的表情基，所述目标表情基的维数为第二维数；

根据包含所述说话人的脸部表情和对应语音的第二媒体数据和所述目标表情基，确定目标表情参数和声学特征；所述目标表情参数用于标识所述说话人说出所述语音的脸部表情相对于所述目标表情基的变化程度；

根据所述目标表情参数和声学特征，驱动具有所述第二表情基的所述第二动画形象。

第四方面，本申请实施例提供一种动画形象驱动装置，所述装置部署在音视频处理设备上，所述装置包括获取单元、第一确定单元、第二确定单元、第三确定单元和驱动单元：

所述获取单元，用于获取包含说话人的脸部表情和对应语音的第一媒体数据；

所述第一确定单元，用于根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；所述第一表情基的维数为第一维数，顶点拓扑为第一顶点拓扑；

所述第二确定单元，用于根据所述第一表情基和待驱动的第二动画形象的第二表情基，确定目标表情基；所述第二表情基的维数为第二维数，顶点拓扑为第二顶点拓扑，所述目标表情基为具有第二顶点拓扑的第一动画形象对应的表情基，所述目标表情基的维数为第二维数；

所述第三确定单元，用于根据包含所述说话人的脸部表情和对应语音的第二媒体数据和所述目标表情基，确定目标表情参数和声学特征；所述目标表情参数用于标识所述说话人说出所述语音的脸部表情相对于所述目标表情基的变化程度；

所述驱动单元，用于根据所述目标表情参数和声学特征，驱动具有所述第二表情基的所述第二动画形象。

第五方面，本申请实施例提供一种用于动画形象驱动的设备，所述设备包括处理器以及存储器：

所述存储器用于存储程序代码，并将所述程序代码传输给所述处理器；

所述处理器用于根据所述程序代码中的指令执行第一方面或第三方面所述的方法。

第六方面，本申请实施例提供一种计算机可读存储介质，所述计算机可读存储介质用于存储程序代码，所述程序代码用于执行第一方面或第三方面所述的方法。

第七方面，本申请实施例提供一种计算机程序产品，当所述计算机程序产品被执行时，用于执行第一方面或第三方面所述的方法。

由上述技术方案可以看出，通过采集说话人说出语音时脸部表情变化的媒体数据，可以确定说话人所对应第一动画形象的第一表情基，第一表情基可以体现第一动画形象的不同表情。在确定出用于驱动第二动画形象的目标文本信息后，可以根据目标文本信息、前述采集的媒体数据和第一表情基，确定对应目标文本信息的声学特征和目标表情参数，该声学特征可以标识模拟所述说话人说出所述目标文本信息的声音，该目标表情参数可以标识模拟所述说话人说出所述目标文本信息的脸部表情相对于所述第一表情基的变化程度。通过声学特征和目标表情参数，可以驱动具有第二表情基的第二动画形象，使得第二动画形象可以通过声学特征模拟发出说话人说出目标文本信息的声音，并且在发声过程中做出符合该说话人应有表情的脸部表情，给用户带来逼真的代入感和沉浸感，提高了用户与动画形象进行交互的体验。

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术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性的前提下，还可以根据这些附图获得其他的附图。

图 1 为本申请实施例提供的一种基于人工智能的动画形象驱动方法的应用场景示意图；

图 2 为本申请实施例提供的一种基于人工智能的动画形象驱动方法的流程图；

图 3 为本申请实施例提供的一种动画形象驱动系统的结构流程；

图 4 为本申请实施例提供的采集媒体数据的场景示例图；

图 5 为本申请实施例提供的 3DMM 库 M 的各个维度分布和意义示例图；

图 6 为本申请实施例提供的一种基于确定捏脸参数的动画形象驱动方法的应用场景示意图；

图 7 为本申请实施例提供的一种基于确定映射关系的动画形象驱动方法的应用场景示意图；

图 8 为本申请实施例提供的时间区间与音素的对应关系的示例图；

图 9 为本申请实施例提供的一种基于人工智能的动画形象驱动方法的流程图；

图 10a 为本申请实施例提供的一种基于人工智能的动画形象驱动方法的流程图；

图 10b 为本申请实施例提供的一种动画形象驱动装置的结构图；

图 11 为本申请实施例提供的一种动画形象驱动装置的结构图；

图 12 为本申请实施例提供的一种用于动画形象驱动的设备的结构图；

图 13 为本申请实施例提供的一种服务器的结构图。

具体实施方式

下面结合附图，对本申请的实施例进行描述。

目前，将具有表情变化能力的动画形象作为与用户交互的交互对象是人机交互的主要研究方向。

例如，在游戏场景中，可以构建和用户自己脸型一样的游戏人物（动画形象）等，当用户输入文字或语音时，该游戏人物可以发出语音并做出对应的表情（例如口型等）；或者，在游戏场景中，构建和用户自己脸型一样的游戏人物等，当对方输入文字或语音时，该游戏人物可以根据对方的输入回复语音并做出对应的表情。

为了可以更好的驱动动画形象，例如驱动动画形象发出语音并做出对应的表情，本申请实施例提供一种基于人工智能的动画形象驱动方法。该方法通过采集说话人说出语音时脸部表情变化的媒体数据，可以确定说话人所对应第一动画形象的第一表情基，在确定出用于驱动第二动画形象的目标文本信息后，可以根据目标文本信息、前述采集的媒体数据和第一表情基，确定对应目标文本信息的声学特征和目标表情参数，从而通过声学特征和

目标表情参数驱动具有第二表情基的第二动画形象，使得第二动画形象通过声学特征模拟发出说话人说出目标文本信息的声音，并且在发声过程中做出符合该说话人应有表情的脸部表情，实现基于文本信息驱动第二动画形象。

需要强调的是，本申请实施例所提供的动画形象驱动方法是基于人工智能实现的，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/深度学习等几大方向。

在本申请实施例中，主要涉及的人工智能技术包括语音处理技术、机器学习和计算机视觉（图像）等方向。

例如可以涉及语音技术(Speech Technology)中的语音识别技术（Automatic Speech Recognition，ASR）、语音合成(Text To Speech，TTS)和声纹识别。语音识别技术中包括语音信号预处理（Speech signal preprocessing）、语音信号频域分析（Speech signal frequency analyzing）、语音信号特征提取（Speech signal feature extraction）、语音信号特征匹配/识别（Speech signal feature matching / recognition）、语音的训练（Speech training）等。语音合成中包括文本分析（Text analyzing）、语音生成（Speech generation）等。

例如可以涉及机器学习（Machine learning，ML），机器学习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。机器学习通常包括深度学习（Deep Learning）等技术，深度学习包括人工神经网络（artificial neural network），例如卷积神经网络(Convolutional Neural Network，CNN)、循环神经网络（Recurrent Neural Network，RNN）、深度神经网络（Deep neural network，DNN）等。

例如可以涉及计算机视觉（Computer Vision）中的视频处理（video processing)、视频语义理解(video semantic understanding，VSU)、人脸识别（face recognition）等。视频语义理解中包括目标识别（target recognition）、目标检测与定位（target detection/localization）等；人脸识别中包括人脸 3D 重建（Face 3D Reconstruction）、人脸检测（Face Detection）、人脸跟踪（Face Tracking）等。

本申请实施例提供的基于人工智能的动画形象驱动方法可以应用于具有驱动动画形象能力的音视频处理设备上，该音视频处理设备可以是终端设备，也可以是服务器。

该音视频处理设备可以具有实施语音技术的能力，让音视频处理设备能听、能看、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。

在本申请实施例中，音视频处理设备通过实施上述计算机视觉技术可以确定媒体数据中说话人所对应第一动画形象的第一表情基，通过语音技术和机器学习可以根据目标文本信息和媒体数据，确定对应目标文本信息的声学特征和目标表情参数，进而利用声学特征和目标表情参数，驱动具有第二表情基的第二动画形象。

其中，若音视频处理设备是终端设备，则终端设备可以是智能终端、计算机、个人数字助理（Personal Digital Assistant，简称 PDA）、平板电脑等。

若该音视频处理设备是服务器，则服务器可以为独立服务器，也可以为集群服务器。当服务器实施该方法时，终端设备可以将包含说话人的脸部表情和对应语音的媒体数据上传给服务器，服务器确定出声学特征和目标表情参数，利用该声学特征和目标表情参数驱动终端设备上的第二动画形象。

可以理解的是，本申请实施例提供的基于人工智能的动画形象驱动方法可以应用到各种适用动画形象的应用场景，例如新闻播报、天气预报、游戏解说以及游戏场景中允许用于构建和用户自己脸型一样的游戏人物等，还能用于利用动画形象承担私人化的服务的场景，例如心理医生，虚拟助手等面向个人的一对一服务。在这些场景下，利用本申请实施例提供的方法可以实现动画形象的驱动。

为了便于理解本申请的技术方案，下面结合实际应用场景对本申请实施例提供的基于人工智能的动画形象驱动方法进行介绍。

参见图 1，图 1 为本申请实施例提供的基于人工智能的动画形象驱动方法的应用场景示意图。该应用场景以音视频处理设备为终端设备为例进行介绍，该应用场景中包括终端设备 101，终端设备 101 可以获取包含说话人的脸部表情和对应语音的媒体数据。该媒体数据可以是一个，也可以是多个。媒体数据可以是视频，也可以是视频和音频。媒体数据中语音包括的字符所对应的语种可以是汉语、英语、韩语等各种语种。

脸部表情可以是说话人说出语音时脸部所做出的动作，例如可以包括口型、眼睛动作、眉毛动作等，视频观看者通过说话人的脸部表情可以感受到媒体数据中的语音就是该说话人说出的。

终端设备 101 根据脸部表情可以确定说话人所对应第一动画形象的第一表情基，第一表情基用于标识第一动画形象的不同表情。

终端设备 101 在确定出用于驱动第二动画形象的目标文本信息后，可以根据目标文本信息、前述采集的媒体数据和第一表情基，确定对应目标文本信息的声学特征和目标表情

参数。其中，表情参数以及后续可能涉及到的捏脸参数的一种表现形式可以是系数，例如可以是具有某一维数的向量。

由于媒体数据中语音和脸部表情是同步的，声学特征和目标表情参数均是根据媒体数据得到的，所对应的是同一个时间轴，故，声学特征所标识的声音和目标表情参数所标识的表情在同一时间轴上同步变化。生成的声学特征是与时间轴相关的一个序列，目标表情参数是与同一时间轴相关的序列，二者可以随着文本信息的变化而有相应的调整。但无论如何调整，声学特征用于标识模拟上述媒体数据中说话人说出目标文本信息的声音，目标表情参数用于标识模拟上述媒体数据中说话人说出目标文本信息的脸部表情相对于所述第一表情基的变化程度。

之后，终端设备 101 可以通过声学特征和目标表情参数，驱动具有第二表情基的第二动画形象，使得第二动画形象可以通过声学特征模拟发出说话人说出目标文本信息的声音，并且在发声过程中做出符合该说话人应有表情的脸部表情。其中，第二动画形象可以是与第一动画形象相同的动画形象，也可以是与第一动画形象不同的动画形象，本申请实施例对此不做限定。

接下来，将结合附图对本申请实施例提供的基于人工智能的动画形象驱动方法进行详细介绍。参见图 2，所述方法包括：

S201、获取包含说话人的脸部表情和对应语音的媒体数据。

包含了脸部表情和对应语音的媒体数据可以是在有摄像头的录音环境下，录制说话人说出的语音，以及通过摄像头录制说话人对应的脸部表情得到的。

若通过摄像头采集到的视频中同时包括说话人的脸部表情和对应语音，则媒体数据为该视频；若通过摄像头采集到的视频中包括说话人的脸部表情，而语音是通过其他设备例如录音设备采集的，则媒体数据包括视频和音频，此时，该视频和音频是同步采集的，视频中包括说话人的脸部表情，音频中包括说话人的语音。

需要说明的是，本申请实施例提供的方法可以通过动画形象驱动系统实现，该系统可以参见图 3 所示，主要包括四个部分，分别是数据采集模块、脸部建模模块、声学特征和表情参数确定模块以及动画驱动模块。其中，数据采集模块用于执行 S201，脸部建模模块用于执行 S202，声学特征和表情参数确定模块用于执行 S203，动画驱动模块用于执行 S204。

包含说话人的脸部表情和对应语音的媒体数据可以是通过数据采集模块得到的。该数据采集模块可以有较多的选择，该数据采集模块可以通过专业设备采集包括说话人的语音和脸部表情的媒体数据，比如使用动作捕捉系统、脸部表情捕捉系统等专业设备来捕捉说话人的脸部表情，脸部表情例如可以是脸部动作、表情、口型等等，使用专业录音设备录制说话人的语音，不同设备之间通过同步信号触发实现语音和脸部表情的数据同步等等。

当然，专业设备并不局限于使用昂贵的捕捉系统，也可以是多视角超高清设备，通过多视角超高清设备采集包括说话人的语音和脸部表情的视频。

该数据采集模块还可以利用多相机环绕的方式采集包括说话人的语音和脸部表情的媒体数据。在一种可能的实现方式中，可以选择 3 个，5 个，乃至更多的超高清相机，正面围绕说话人拍摄。采集环境中需要有稳定的环境光照，不要求说话人穿特定的衣服。参见图 4 所示，图 4 以 3 个超高清相机为例，上方虚线箭头表示稳定光照，左侧三个箭头表示超高清相机的视角和说话人的关系，从而采集包括说话人的语音和脸部表情的媒体数据。此时，通过超高清相机采集的视频中可以同时包括语音和脸部表情，即媒体数据为视频。

需要说明的是，在采集媒体数据时，根据采集脸部表情所使用传感器的不同，采集的媒体数据的表现形式可以有所不同。在一些情况下，可以通过具有红绿蓝深度（Red Green Blue Deep，RGBD）传感器的对说话人进行拍摄，实现对脸部模型的建立。由于 RGBD 传感器可以采集到深度信息，得到说话人的三维重建结果，因此，媒体数据中包括说话人对应的脸部静态建模，即 3 维（3 Dimensions，3D）数据。在另外一些情况下，可能没有 RGBD 传感器而是使用二维传感器对说话人进行拍摄，此时，没有说话人的三维重建结果，媒体数据中包括说话人对应的视频帧，即 2 维（2 Dimensions，2D）数据。

S202、根据脸部表情确定该说话人所对应第一动画形象的第一表情基。

在获取到上述媒体数据后，通过图 3 中的脸部建模模块可以对说话人进行脸部建模，从而得到说话人所对应第一动画形象的第一表情基，该第一表情基用于标识所述第一动画形象的表情。

进行脸部建模的目的在于使得被采集的对象例如前述提到的说话人可以被计算机理解并存储，包括被采集对象的形状、纹理等。进行脸部建模的方式可以包括多种，主要从硬件、人工、软件三个角度来实现。其中，硬件角度实现可以是采用专业设备对说话人进行高精度的扫描，如 3D 扫描仪器，对得到的脸部模型可以选择手动/自动清理数据；人工角度实现可以是由美术设计师手工设计数据、清理数据、调节数据；软件角度实现可以是采用参数化捏脸算法自动生成说话人脸部模型。

在表情参数化时，同样可以从硬件、人工、软件三个角度来实现。比如可以使用专业人脸扫描设备扫描带有表情的说话人之后，会自动给出对当前表情的参数化描述，这种描述与扫描设备中自定义的表情描述相关。而对于美术设计师手工调节的表情参数，一般需要预先定义表情类型和对应的人脸参数化，比如嘴巴的张合程度，脸部肌肉的运动幅度等等。而对于软件实现表情参数化，一般需要定义脸部在不同表情中的数学描述，比如通过对大量真实脸部数据，进行主成分分析方法 (Principal Component Analysis，PCA)分解之后，得到最能体现各个表情相对平均脸的变化程度的数字描述。

在本实施例中，主要对基于软件的脸部建模和表情参数化进行介绍。在这种情况下，脸部在不同表情中的数学描述可以通过模型库定义。本申请实施例中的动画形象（例如第一动画形象和后续的第二动画形象）可以为模型库中的模型，也可以是通过模型库中模型的线性组合得到的。该模型库可以是人脸 3D 可变形模型 (3DMM) 库，也可以是其他模型库，本实施对此不做限定。动画形象可以是一个 3D 网格。

以 3DMM 库为例，3DMM 库由大量高精度脸部数据通过主成分分析方法得到，描述了高维脸型和表情相对平均脸的主要变化，也可以描述纹理信息。

一般来说，3DMM 库描述一个无表情的脸型时，可以通过$\mathrm{mu} + \sum(\mathrm{Pface_i} - \mathrm{mu}) * \alpha_i$ 得到。其中，mu 是自然表情下的平均脸，$Pface_i$是第 i 个脸型主成分分量，α_i就是各个脸型主成分分量的权重，也就是捏脸参数。

假设 3DMM 库中的动画形象对应的网格可以通过M表示，即通过 M 表示 3DMM 库中的脸型、表情和顶点之间的关系，M是一个$[m \times n \times d]$的三维矩阵，其中每一维分别为网格的顶点坐标(m)、脸型主成分(n)、表情主成分(d)。3DMM 库 M 的各个维度分布和意义如图 5 所示，每个坐标轴分别表示顶点坐标(m)、脸型主成分(n)、表情主成分(d)。由于 m 表示 xyz 三个坐标的值，所以网格的顶点数为$m/3$，记作 v。如果确定了动画形象的脸型或者表情，那么 M 可以是一个二维矩阵。

在本申请实施例中，不考虑 3DMM 库中的纹理维度，假设动画形象的驱动为F，则：

$$F = M \times \alpha \times \beta = \sum_{i=1}^{d}(\sum_{j=1}^{n} M_{k,j,i} * \alpha_j) * \beta_i \qquad (1)$$

其中，M 为动画形象的网格，α为捏脸参数，β为表情参数；n 为捏脸基中捏脸网格的个数，d 为表情基中表情网格的个数，$M_{k,j,i}$为具有第 i 个表情网格、第 j 个捏脸网格的第 k 个网格，α_j为一组捏脸参数中的第 j 维，表示第 j 个脸型主成分分量的权重，β_i为一组表情参数中的第 i 维，表示第 i 个表情主成分分量的权重。

其中，确定捏脸参数的过程为捏脸算法，确定表情参数的过程为捏表情算法。捏脸参数用于与捏脸基做线性组合得到对应的脸型，例如存在一个包括 50 个捏脸网格（属于可变形网格，例如 blendshape）的捏脸基，该捏脸基对应的捏脸参数为一个 50 维的向量，每一维可以标识该捏脸参数所对应脸型与一个捏脸网格的相关程度。捏脸基所包括的捏脸网格分别代表不同脸型，每一个捏脸网格均为相对平均脸变化较大的脸部形象，是大量的脸通过 PCA 分解之后的得到的不同维度的脸型主成分，且同一个捏脸基中不同捏脸网格对应的顶点序号保持一致。

表情参数用于与表情基做线性组合得到对应的表情，例如存在一个包括 50 个（相当于维数为 50）表情网格（属于可变形网格，例如 blendshape）的表情基，该表情基对应的表情参数为一个 50 维的向量，每一维可以标识该表情参数所对应表情与一个表情网格的相关程度。表情基所包括的表情网格分别代表不同表情，每一个表情网格均由同一个 3D 模型在不同表情下变化而成，同一个表情基中不同表情网格对应的顶点序号保持一致。

针对前述的可变形网格，单个网格可以通过预定义形状变形，得到任意数量网格。

结合上述公式（1），可以得到说话人所对应第一动画形象的第一表情基，从而用于后续第二动画形象的驱动。

S203、根据目标文本信息、该媒体数据和第一表情基确定对应目标文本信息的声学特征和目标表情参数。

通过图 3 中的声学特征和表情参数确定模块可以确定对应目标文本信息的声学特征和目标表情参数。其中，声学特征用于标识模拟说话人说出目标文本信息的声音，目标表情参数用于标识模拟说话人说出目标文本信息的脸部表情相对于第一表情基的变化程度。

可以理解的是，目标文本信息的获取方式可以包括多种，例如，目标文本信息可以是用户通过终端设备输入的，也可以是根据输入至终端设备的语音转换得到的。

S204、根据声学特征和目标表情参数，驱动具有第二表情基的第二动画形象。

通过图 3 中的动画驱动模块将目标表情参数所标识的表情，配合声学特征所标识的语音，通过人类能直观理解的方式，利用多种感官来展示。其中一种可行的方式是，假设目标表情参数表示了第二表情基中各个表情网格的权重，通过第二表情基加权线性组合可以得到对应的表情。在发出语音的同时，通过渲染方法将做出与该语音对应表情的第二动画形象渲染出来，从而实现第二动画形象的驱动。

由上述技术方案可以看出，通过采集说话人说出语音时脸部表情变化的视频，可以确定说话人所对应第一动画形象的第一表情基，第一表情基可以体现第一动画形象的不同表情。在确定出用于驱动第二动画形象的目标文本信息后，可以根据目标文本信息、前述采集的媒体数据和第一表情基，确定对应目标文本信息的声学特征和目标表情参数，该声学特征可以标识模拟所述说话人说出所述目标文本信息的声音，该目标表情参数可以标识模拟所述说话人说出所述目标文本信息的脸部表情相对于所述第一表情基的变化程度。通过声学特征和目标表情参数，可以驱动具有第二表情基的第二动画形象，使得第二动画形象可以通过声学特征模拟发出说话人说出目标文本信息的声音，并且在发声过程中做出符合该说话人应有表情的脸部表情，给用户带来逼真的代入感和沉浸感，提高了用户与动画形象进行交互的体验。

需要说明的是，S203 的实现方式可以包括多种，本申请实施例着重对一种实现方式进行介绍。

在一种可能的实现方式中，S203 的实现方式可以是根据目标文本信息和媒体数据，确定对应目标文本信息的声学特征和表情特征。该声学特征用于标识模拟说话人说出目标文本信息的声音，该表情特征用于标识模拟说话人说出所述目标文本信息的脸部表情。然后，根据第一表情基和表情特征确定目标表情参数。

由于媒体数据中已经同步记录了说话人的脸部表情和语音，即媒体数据中说话人的脸部表情和语音对应同一时间轴。故，可以在线下预先收集大量的媒体数据作为训练数据，从这些媒体数据中提取文本特征、声学特征和表情特征，根据这些特征训练得到时长模型、声学模型、表情模型。当线上获取包含说话人的脸部表情和对应语音的媒体数据时，可以使用时长模型确定目标文本信息对应的时长，再将时长结合目标文本信息对应的文本特征

分别通过声学模型和表情模型确定对应的声学特征和表情特征。由于声学特征和表情特征都是基于同一个时长模型得到的时长，因此很容易做到语音以及表情的同步，从而使得第二动画形象在模拟说话人说出目标文本信息对应语音的同时，模拟说话人做出相应的表情。

接下来，将对 S204 的可能实现方式进行介绍。应理解，在本实施例中，第二动画形象可以是与第一动画形象相同的动画形象，也可以是与第一动画形象不同的动画形象。在这两种情况下，S204 的实现方式可能有所不同。

第一种情况：第一动画形象和第二动画形象为同一个动画形象。

在这种情况下，所需驱动的动画形象即第一动画形象。那么，为了驱动第一动画形象，除了需要确定第一表情基，还需要确定第一动画形象的捏脸参数，得到第一动画形象的脸型。因此，在 S202 中，可以根据脸部表情确定第一动画形象的第一表情基和第一动画形象的捏脸参数，该捏脸参数用于标识第一动画形象的脸型相对于第一动画形象所对应捏脸基的变化程度。

确定第一动画形象的第一表情基和第一动画形象的捏脸参数的方式有很多种。在一些情况下，基于媒体数据确定捏脸参数以建立脸部模型时，所采集的媒体数据往往精度不高，噪声较大，使得建立的脸部模型质量不高，具有很多不确定性，难以准确体现待建对象的实际外形。例如，由于采集不规范导致建模质量低；重建过程容易受到环境光照、用户化妆等影响；重建的脸部模型中含有表情，并非自然状态；建立的脸部模型无法适应之后将要提取表情参数的视频等。为了解决这一问题，本申请实施例提供一种捏脸参数的确定方法，参见图 6 所示。

在图 6 中，若获取的媒体数据中可以包括多组脸部顶点数据，可以基于其中的第一顶点数据，以及 3DMM 库中用于标识目标脸部模型的目标顶点数据，确定初始捏脸参数。在确定出初始捏脸参数的基础上，通过获取媒体数据中的第二顶点数据，基于初始捏脸参数和目标顶点数据确定表情参数，之后，固定该表情参数，反推捏脸参数或者说反推如何变化脸型得到在该表情参数下的说话人的脸部形象，即通过固定表情反推脸型的方式修正初始捏脸参数，得到目标捏脸参数，从而将该目标捏脸参数作为第一动画形象的捏脸参数。

由于第二顶点数据和第一顶点数据分别标识待建对象的不同脸部形象，故第二顶点数据和第一顶点数据受到完全相同的不确定性影响的几率较小，在通过第一顶点数据确定出初始捏脸参数的基础上，通过第二顶点数据修正出的目标捏脸参数可以一定程度上抵消第一顶点数据中的噪声，以目标捏脸参数确定出的说话人对应的脸部模型精确度相对更高。

由于第一表情基与第二表情基相同，即二者的维数以及各个维数的语义信息相同，确定出的目标表情参数可以直接驱动第二动画形象，故，在 S204 中驱动第二动画形象的方式可以是根据声学特征、目标表情参数和捏脸参数，驱动具有第二表情基的第二动画形象。

第二种情况：第一动画形象和第二动画形象为不同动画形象。

在这种情况下， 第一表情基与第二表情基不同，即二者的维数以及各个维数的语义信息存在不同，故难以直接利用目标表情参数驱动具有第二表情基的第二动画形象。由于第一动画形象对应的表情参数与第二动画形象对应的表情参数应具有映射关系，第一动画形象对应的表情参数与第二动画形象对应的表情参数间的映射关系可以通过函数 f()表示，则通过第一动画形象对应的表情参数计算第二动画形象对应的表情参数的公式如下：

$$\beta^b = f(\beta^a) \qquad (2)$$

其中，β^b为第二动画形象对应的表情参数，β^a为第一动画形象对应的表情参数，f()表示第一动画形象对应的表情参数与第二动画形象对应的表情参数间的映射关系。

故，若确定出该映射关系，便可以利用第一动画形象（例如动画形象 a）对应的表情参数直接驱动第二动画形象（例如动画形象 b）。其中，映射关系可以是线性映射关系，也可以是非线性映射关系。

为了实现根据目标表情参数驱动具有第二表情基的第二动画形象，需要确定出映射关系。确定映射关系的方式可以包括多种，本实施例主要对两种确定方式进行介绍。

第一种确定方式可以是基于第一动画形象对应的第一表情基和第二动画形象对应的第二表情基，确定表情参数间的映射关系。参见图 7 所示，由于第一动画形象对应的实际表情参数可以驱动被第一动画形象做出实际表情，该实际表情参数可以体现该实际表情与其表情基的不同维度下的相关程度，即第二动画形象对应的实际表情参数也可以体现第二动画形象的实际表情与其表情基的不同维度下的相关程度，故基于上述表情参数与表情基间的关联关系，可以根据第一动画形象对应的第一表情基和第二动画形象对应的第二表情基，确定出表情参数间的映射关系。然后，根据声学特征、目标表情参数和该映射关系，驱动具有第二表情基的第二动画形象。

第二种确定方式可以是基于音素和第二表情基之间的预设关系确定表情参数间的映射关系。

音素是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（例如口型）构成一个音素。也就是说，音素与说话人无关，无论说话人是谁、无论语音是英语还是汉语、无论发出音素所对应的文本是否相同，只要语音中一个时间区间内的音素相同，那么，对应的表情例如口型具有一致性。参见图 8 所示，图 8 示出了时间区间与音素的对应关系，描述了在一个语音中，哪个时间区间对应了哪个音素。例如，第二行中“5650000”和“6300000”代表时间戳，表示 5.65 秒至 6.3 秒这一时间区间，在该时间区间内说话人发出的音素是“u”。音素的统计方法并不唯一，本实施例以 33 个中文音素为例。

由于媒体数据中，面部表情和语音是同步采集的，因此可以方便的通过语音的划分，得到对应的视频帧，即根据媒体数据确定语音所标识音素、该音素对应的时间区间和媒体

数据处于该时间区间的视频帧。然后，根据该视频帧确定音素对应的第一表情参数，第一表情参数用于标识发出该音素时说话人的脸部表情相对于第一表情基的变化程度。

例如图 8 中第二行，对于音素“u”，其所对应的时间区间是 5.65 秒至 6.3 秒，确定处于时间区间 5.65 秒至 6.3 秒的视频帧，根据该视频帧提取音素“u”对应的第一表情参数。若第一动画形象为动画形象 a，第一表情参数可以用β^a表示。若第一表情基的维数是n^a，则得到的第一表情参数β^a为一组n^a长度的向量。

由于该确定映射关系的方式的前提是其他动画形象的表情基例如第二动画形象对应的第二表情基是根据与音素的预设关系生成的，预设关系表示的是一个音素对应一个表情网格，比如对于第二动画形象 b 而言，预设关系中音素“u”对应第 1 个表情网格，音素“i”对应第 2 个表情网格……，若音素的个数为n^b个，则根据预设关系可以确定出包括n^b个表情网格的第二表情基。那么，当确定出语音所标识的音素后，便可以根据预设关系和第二表情基，确定该音素对应的第二表情参数。然后，根据第一表情参数和第二表情参数，确定映射关系。

例如，语音所标识的音素为“u”，通过第二表情基和预设关系可知音素“u”对应第 1 个表情网格，则可以确定出第二表情参数为$\beta^b = [1\ 0\ \ldots\ 0]$，$\beta^b$中包括$n^b$个元素，除了第一个元素为 1，其余$n^b - 1$个元素均为 0。

由此，一组β^b和β^a的映射关系就建立了。当得到大量第一表情参数β^a时，可以产生大量对应的第二表情参数β^b。假设第一表情参数β^a和第二表情参数β^b的个数分别是 L 个，L 个第一表情参数β^a构成第一矩阵，L 个第二表情参数β^b构成第二矩阵，分别记作β^A和β^B。有：

$$\beta^A = [L \times n^a],\quad \beta^B = [L \times n^b] \qquad (3)$$

本方案以第一表情参数和第二表情参数之间满足线性映射关系为例，则上述公式（2）可以变形为：

$$\beta^b = f * \beta^a \qquad (4)$$

根据公式（3）和（4）所示，则映射关系的确定公式可以为：

$$f = \beta^B * inv(\beta^A) \qquad (5)$$

其中，f 为映射关系，β^A为第一矩阵，β^B为第二矩阵，inv为矩阵求逆运算。

在得到映射关系 f 后，对于任意一组第一表情参数β^a，可以得到对应的$\beta^b = f * \beta^a$，从而根据第一表情参数得到第二表情参数，以便驱动第二动画形象，例如动画形象 b。

前述实施例主要介绍了如何基于文本信息驱动动画形象。在一些情况下，还可以基于媒体数据直接驱动动画形象。例如，媒体数据中说话人所对应第一动画形象具有第一表情基，第一表情基的维数为第一维数，顶点拓扑为第一顶点拓扑，第一表情基可以用 Ea 表示，

第一维数可以用 Na 表示，第一顶点拓扑可以用 Ta 表示，第一表情基 Ea 的样子是 Fa；待驱动的第二动画形象具有第二表情基，第二表情基的维数为第二维数，顶点拓扑为第二顶点拓扑，第二表情基可以用 Eb 表示，第二维数可以用 Nb 表示，第二顶点拓扑可以用 Tb 表示，第二表情基 Eb 的样子是 Fb，希望通过包括该说话人脸部表情和语音的媒体数据来驱动第二动画形象。

为此，本申请实施例还提供一种基于人工智能的动画形象驱动方法，参见图 9 所示，所述方法包括：

S901、获取包含说话人的脸部表情和对应语音的第一媒体数据。

S902、根据脸部表情确定所述说话人所对应第一动画形象的第一表情基。

S903、根据第一表情基和待驱动的第二动画形象的第二表情基，确定目标表情基。

在本实施例中，由于第一表情基的维数与第二表情基的维数不同，为了可以利用媒体数据中该说话人的脸部表情和语音驱动第二动画形象，可以构造出一个新的表情基例如目标表情基，使得该目标表情基同时具有第一表情基和第二表情基的特点。

在一种实现方式中，S903 的实现方式可以是：从第一表情基中确定第一动画形象处于无表情时对应的无表情网格，并从第二表情基中确第二动画形象处于无表情时对应的无表情网格。根据第一形象对应的无表情网格和第二形象对应的无表情网格，确定调整网格，该调整网格具有第二顶点拓扑，用于标识处于无表情时的第一动画形象。根据调整网格和第二表情基中的网格形变关系，生成目标表情基。

若第一表情基为 Ea，第一维数为 Na，第一顶点拓扑为 Ta，第一表情基 Ea 的样子是 Fa；第二表情基为 Eb，第二维数为 Nb，第二顶点拓扑为 Tb，第二表情基 Eb 的样子是 Fb，则该方法的流程图还可以参见图 10a 所示。基于第一表情基 Ea 和第二表情基 Eb 确定目标表情基 Eb’。其中，确定目标表情基 Eb’的方式可以是提取第二表情基 Eb 的无表情网格和第一表情基 Ea 的无表情网格。通过捏脸算法例如 nricp 算法，将 Eb 的无表情网格贴到 Ea 的无表情网格上，使得 Eb 的无表情网格在保持顶点拓扑 Fb 的前提下，改变样子，变成 Ea 的样子，得到调整网格，该调整网格可以表示为 Newb。随后，由于 Newb 和第二表情基 Eb 中各个维度的表情相对自然表情（无表情）的网格形变关系是已知的，故，可以根据 Newb 和第二表情基 Eb 中的网格形变关系从 Newb 中形变出目标表情基 Eb’。目标表情基 Eb’的样子是 Fa，维数是 Nb，顶点拓扑是 Tb。

S904、根据包含所述说话人的脸部表情和对应语音的第二媒体数据和所述目标表情基，确定目标表情参数和声学特征。

在得到目标表情基后，参见图 10a 所示，基于根据包含该说话人的脸部表情和对应语音的第二媒体数据和该目标表情基 Eb’，提取出声学特征并通过捏表情算法得到目标表情参数 Bb。其中，目标表情参数用于标识说话人说出所述语音的脸部表情相对于目标表情基的变化程度。

可以理解的是，利用该方法得到的目标表情参数和声学特征可以用于重新训练前述所提到的声学模型、表情模型。

S905、根据目标表情参数和声学特征，驱动具有第二表情基的所述第二动画形象。

S901、S902 和 S905 的具体实现方式分别可以参见前述 S201、S202 和 S204 的实现方式，此处不再赘述。

接下来，将结合实际应用场景对本申请实施例提供的基于人工智能的动画形象驱动方法进行介绍。

在该应用场景中，在该应用场景中，假设第一动画形象为仿照说话人的形象构建的，第二动画形象为在游戏中与用户进行交互的游戏角色的形象。当该游戏角色通过输入的目标文本信息与用户进行交流时，希望通过该目标文本信息驱动该游戏角色模仿说话人发出目标文本信息对应的语音，并做出对应的表情。故，终端设备可以获取包含说话人的脸部表情和对应语音的媒体数据，根据脸部表情确定该说话人所对应第一动画形象的第一表情基。接着，根据目标文本信息、媒体数据和第一表情基，确定对应所述目标文本信息的声学特征和目标表情参数，从而根据该声学特征和目标表情参数，驱动具有第二表情基的第二动画形象，使得第二动画形象发出目标文本信息对应的语音，并做出对应的表情。这样，用户可以看到该游戏角色模仿说话人说出语音，并做出对应的表情，为用户带来逼真的代入感和沉浸感，提高了用户与动画形象进行交互的体验。

基于前述实施例提供的方法，本实施例还提供一种动画形象驱动装置 1000，所述装置 1000 部署在音视频处理设备上。参见图 10b，所述装置 1000 包括获取单元 1001、第一确定单元 1002、第二确定单元 1003 和驱动单元 1004：

所述获取单元 1001，用于获取包含说话人的脸部表情和对应语音的媒体数据；

所述第一确定单元 1002，用于根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；

所述第二确定单元 1003，用于根据目标文本信息、所述媒体数据和所述第一表情基，确定对应所述目标文本信息的声学特征和目标表情参数；所述声学特征用于标识模拟所述说话人说出所述目标文本信息的声音，所述目标表情参数用于标识模拟所述说话人说出所述目标文本信息的脸部表情相对于所述第一表情基的变化程度；

所述驱动单元 1004，用于根据所述声学特征和所述目标表情参数，驱动具有第二表情基的第二动画形象。

在一种可能的实现方式中，所述第一动画形象和所述第二动画形象为同一个动画形象，所述第一表情基与所述第二表情基相同，所述第一确定单元 1002，用于：

根据所述脸部表情确定所述第一动画形象的第一表情基和所述第一动画形象的捏脸参数，所述捏脸参数用于标识所述第一动画形象的脸型相对于所述第一动画形象所对应捏脸基的变化程度；

所述驱动单元 1004，用于：

根据所述声学特征、所述目标表情参数和所述捏脸参数，驱动所述第二动画形象。

在一种可能的实现方式中，所述第一动画形象和所述第二动画形象为不同动画形象，所述第一表情基与所述第二表情基不同，所述驱动单元 1004，用于：

确定所述第一表情基所对应表情参数与所述第二表情基所对应表情参数间的映射关系；

根据所述声学特征、所述目标表情参数和所述映射关系，驱动所述第二动画形象。

在一种可能的实现方式中，所述第二表情基是根据所述第二表情基与音素的预设关系生成的，所述驱动单元 1004，还用于：

根据所述媒体数据确定所述语音所标识音素、所述音素对应的时间区间和所述媒体数据处于所述时间区间的视频帧；

根据所述视频帧确定所述音素对应的第一表情参数，所述第一表情参数用于标识发出所述音素时所述说话人的脸部表情相对于所述第一表情基的变化程度；

根据所述预设关系和所述第二表情基，确定所述音素对应的第二表情参数；

根据所述第一表情参数和所述第二表情参数，确定所述映射关系。

在一种可能的实现方式中，所述第二确定单元 1003，用于：

根据所述目标文本信息和所述媒体数据，确定对应所述目标文本信息的声学特征和表情特征；所述表情特征用于标识模拟所述说话人说出所述目标文本信息的脸部表情；

根据所述第一表情基和所述表情特征确定所述目标表情参数。

本实施例还提供一种动画形象驱动装置 1100，所述装置 1100 部署在音视频处理设备上。参见图 11，所述装置 1100 包括获取单元 1101、第一确定单元 1102、第二确定单元 1103、第三确定单元 1104 和驱动单元 1105：

所述获取单元 1101，用于获取包含说话人的脸部表情和对应语音的第一媒体数据；

所述第一确定单元 1102，用于根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；所述第一表情基的维数为第一维数，顶点拓扑为第一顶点拓扑；

所述第二确定单元 1103，用于根据所述第一表情基和待驱动的第二动画形象的第二表情基，确定目标表情基；所述第二表情基的维数为第二维数，顶点拓扑为第二顶点拓扑，所述目标表情基为具有第二顶点拓扑的第一动画形象对应的表情基，所述目标表情基的维数为第二维数；

所述第三确定单元 1104，用于根据包含所述说话人的脸部表情和对应语音的第二媒体数据和所述目标表情基，确定目标表情参数和声学特征；所述目标表情参数用于标识所述说话人说出所述语音的脸部表情相对于所述目标表情基的变化程度；

所述驱动单元 1105，用于根据所述目标表情参数和声学特征，驱动具有所述第二表情基的所述第二动画形象。

在一种可能的实现方式中，所述第二确定单元 1103，用于从所述第一表情基中确定所述第一动画形象处于无表情时对应的无表情网格，并从所述第二表情基中确所述第二动画形象处于无表情时对应的无表情网格；

根据所述第一动画形象对应的无表情网格和所述第二动画形象对应的无表情网格，确定调整网格，所述调整网格具有第二顶点拓扑，用于标识处于无表情时的第一动画形象；

根据所述调整网格和所述第二表情基中的网格形变关系，生成所述目标表情基。

本申请实施例还提供了一种用于动画形象驱动的设备，该设备可以通过语音驱动动画，该设备可以为音视频处理设备。下面结合附图对该设备进行介绍。请参见图 12 所示，本申请实施例提供了一种用于动画形象驱动的设备，该设还可以是终端设备，该终端设备可以为包括手机、平板电脑、个人数字助理（Personal Digital Assistant，简称 PDA）、销售终端（Point of Sales，简称 POS）、车载电脑等任意智能终端，以终端设备为手机为例：

图 12 示出的是与本申请实施例提供的终端设备相关的手机的部分结构的框图。参考图 12，手机包括：射频（Radio Frequency，简称 RF）电路 1210、存储器 1220、输入单元 1230、显示单元 1240、传感器 1250、音频电路 1260、无线保真（wireless fidelity，简称 WiFi）模块 1270、处理器 1280、以及电源 1290 等部件。本领域技术人员可以理解，图 12 中示出的手机结构并不构成对手机的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

下面结合图 12 对手机的各个构成部件进行具体的介绍：

RF 电路 1210 可用于收发信息或通话过程中，信号的接收和发送，特别地，将基站的下行信息接收后，给处理器 1280 处理；另外，将设计上行的数据发送给基站。通常，RF 电路 1210 包括但不限于天线、至少一个放大器、收发信机、耦合器、低噪声放大器（Low Noise Amplifier，简称 LNA）、双工器等。此外，RF 电路 1210 还可以通过无线通信与网络和其他设备通信。上述无线通信可以使用任一通信标准或协议，包括但不限于全球移动通讯系统（Global System of Mobile communication，简称 GSM）、通用分组无线服务（General Packet Radio Service，简称 GPRS）、码分多址（Code Division Multiple Access，简称 CDMA）、

宽带码分多址（Wideband Code Division Multiple Access，简称 WCDMA）、长期演进（Long Term Evolution，简称 LTE）、电子邮件、短消息服务（Short Messaging Service，简称 SMS）等。

存储器 1220 可用于存储软件程序以及模块，处理器 1280 通过运行存储在存储器 1220 的软件程序以及模块，从而执行手机的各种功能应用以及数据处理。存储器 1220 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（比如声音播放功能、图像播放功能等）等；存储数据区可存储根据手机的使用所创建的数据（比如音频数据、电话本等）等。此外，存储器 1220 可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

输入单元 1230 可用于接收输入的数字或字符信息，以及产生与手机的用户设置以及功能控制有关的键信号输入。具体地，输入单元 1230 可包括触控面板 1231 以及其他输入设备 1232。触控面板 1231，也称为触摸屏，可收集用户在其上或附近的触摸操作（比如用户使用手指、触笔等任何适合的物体或附件在触控面板 1231 上或在触控面板 1231 附近的操作），并根据预先设定的程式驱动相应的连接装置。可选的，触控面板 1231 可包括触摸检测装置和触摸控制器两个部分。其中，触摸检测装置检测用户的触摸方位，并检测触摸操作带来的信号，将信号传送给触摸控制器；触摸控制器从触摸检测装置上接收触摸信息，并将它转换成触点坐标，再送给处理器 1280，并能接收处理器 1280 发来的命令并加以执行。此外，可以采用电阻式、电容式、红外线以及表面声波等多种类型实现触控面板 1231。除了触控面板 1231，输入单元 1230 还可以包括其他输入设备 1232。具体地，其他输入设备 1232 可以包括但不限于物理键盘、功能键（比如音量控制按键、开关按键等）、轨迹球、鼠标、操作杆等中的一种或多种。

显示单元 1240 可用于显示由用户输入的信息或提供给用户的信息以及手机的各种菜单。显示单元 1240 可包括显示面板 1241，可选的，可以采用液晶显示器（Liquid Crystal Display，简称 LCD）、有机发光二极管（Organic Light-Emitting Diode，简称 OLED）等形式来配置显示面板 1241。进一步的，触控面板 1231 可覆盖显示面板 1241，当触控面板 1231 检测到在其上或附近的触摸操作后，传送给处理器 1280 以确定触摸事件的类型，随后处理器 1280 根据触摸事件的类型在显示面板 1241 上提供相应的视觉输出。虽然在图 12 中，触控面板 1231 与显示面板 1241 是作为两个独立的部件来实现手机的输入和输入功能，但是在某些实施例中，可以将触控面板 1231 与显示面板 1241 集成而实现手机的输入和输出功能。

手机还可包括至少一种传感器 1250，比如光传感器、运动传感器以及其他传感器。具体地，光传感器可包括环境光传感器及接近传感器，其中，环境光传感器可根据环境光线的明暗来调节显示面板 1241 的亮度，接近传感器可在手机移动到耳边时，关闭显示面板 1241 和/或背光。作为运动传感器的一种，加速计传感器可检测各个方向上（一般为三轴）加速度的大小，静止时可检测出重力的大小及方向，可用于识别手机姿态的应用（比如横

竖屏切换、相关游戏、磁力计姿态校准）、振动识别相关功能（比如计步器、敲击）等；至于手机还可配置的陀螺仪、气压计、湿度计、温度计、红外线传感器等其他传感器，在此不再赘述。

音频电路 1260、扬声器 1261，传声器 1262 可提供用户与手机之间的音频接口。音频电路 1260 可将接收到的音频数据转换后的电信号，传输到扬声器 1261，由扬声器 1261 转换为声音信号输出；另一方面，传声器 1262 将收集的声音信号转换为电信号，由音频电路 1260 接收后转换为音频数据，再将音频数据输出处理器 1280 处理后，经 RF 电路 1210 以发送给比如另一手机，或者将音频数据输出至存储器 1220 以便进一步处理。

WiFi 属于短距离无线传输技术，手机通过 WiFi 模块 1270 可以帮助用户收发电子邮件、浏览网页和访问流式媒体等，它为用户提供了无线的宽带互联网访问。虽然图 12 示出了 WiFi 模块 1270，但是可以理解的是，其并不属于手机的必须构成，完全可以根据需要在不改变发明的本质的范围内而省略。

处理器 1280 是手机的控制中心，利用各种接口和线路连接整个手机的各个部分，通过运行或执行存储在存储器 1220 内的软件程序和/或模块，以及调用存储在存储器 1220 内的数据，执行手机的各种功能和处理数据，从而对手机进行整体监控。可选的，处理器 1280 可包括一个或多个处理单元；优选的，处理器 1280 可集成应用处理器和调制解调处理器，其中，应用处理器主要处理操作系统、用户界面和应用程序等，调制解调处理器主要处理无线通信。可以理解的是，上述调制解调处理器也可以不集成到处理器 1280 中。

手机还包括给各个部件供电的电源 1290（比如电池），优选的，电源可以通过电源管理系统与处理器 1280 逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。

尽管未示出，手机还可以包括摄像头、蓝牙模块等，在此不再赘述。

在本实施例中，该终端设备所包括的处理器 1280 还具有以下功能：

获取包含说话人的脸部表情和对应语音的媒体数据；

根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；

根据目标文本信息、所述媒体数据和所述第一表情基，确定对应所述目标文本信息的声学特征和目标表情参数；所述声学特征用于标识模拟所述说话人说出所述目标文本信息的声音，所述目标表情参数用于标识模拟所述说话人说出所述目标文本信息的脸部表情相对于所述第一表情基的变化程度；

根据所述声学特征和所述目标表情参数，驱动具有第二表情基的第二动画形象。

或，

获取包含说话人的脸部表情和对应语音的第一媒体数据；

根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；所述第一表情基的维数为第一维数，顶点拓扑为第一顶点拓扑；

根据所述第一表情基和待驱动的第二动画形象的第二表情基，确定目标表情基；所述第二表情基的维数为第二维数，顶点拓扑为第二顶点拓扑，所述目标表情基为具有第二顶点拓扑的第一动画形象对应的表情基，所述目标表情基的维数为第二维数；

根据包含所述说话人的脸部表情和对应语音的第二媒体数据和所述目标表情基，确定目标表情参数和声学特征；所述目标表情参数用于标识所述说话人说出所述语音的脸部表情相对于所述目标表情基的变化程度；

根据所述目标表情参数和声学特征，驱动具有所述第二表情基的所述第二动画形象。

本申请实施例还提供服务器，请参见图 13 所示，图 13 为本申请实施例提供的服务器 1300 的结构图，服务器 1300 可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上中央处理器（Central Processing Units，简称 CPU）1322（例如，一个或一个以上处理器）和存储器 1332，一个或一个以上存储应用程序 1342 或数据 1344 的存储介质 1330（例如一个或一个以上海量存储设备）。其中，存储器 1332 和存储介质 1330 可以是短暂存储或持久存储。存储在存储介质 1330 的程序可以包括一个或一个以上模块（图示没标出），每个模块可以包括对服务器中的一系列指令操作。更进一步地，中央处理器 1322 可以设置为与存储介质 1330 通信，在服务器 1300 上执行存储介质 1330 中的一系列指令操作。

服务器 1300 还可以包括一个或一个以上电源 1326，一个或一个以上有线或无线网络接口 1350，一个或一个以上输入输出接口 1358，和/或，一个或一个以上操作系统 1341，例如 Windows ServerTM，Mac OS XTM，UnixTM, LinuxTM，FreeBSDTM 等等。

上述实施例中由服务器所执行的步骤可以基于该图 13 所示的服务器结构。

本申请实施例还提供一种计算机可读存储介质，所述计算机可读存储介质用于存储程序代码，所述程序代码用于执行前述各个实施例所述的动画形象驱动方法。

本申请实施例还提供一种包括指令的计算机程序产品，当其在计算机上运行时，使得计算机执行前述各个实施例所述的动画形象驱动方法。

本申请的说明书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等（如果存在）是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例例如能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、

产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

应当理解，在本申请中，“至少一个（项）”是指一个或者多个，“多个”是指两个或两个以上。“和/或”，用于描述关联对象的关联关系，表示可以存在三种关系，例如，“A 和/或 B”可以表示：只存在 A，只存在 B 以及同时存在 A 和 B 三种情况，其中 A，B 可以是单数或者复数。字符“/”一般表示前后关联对象是一种“或”的关系。“以下至少一项(个)”或其类似表达，是指这些项中的任意组合，包括单项（个）或复数项（个）的任意组合。例如，a，b 或 c 中的至少一项（个），可以表示：a，b，c，“a 和 b”，“a 和 c”，“b 和 c”，或“a 和 b 和 c”，其中 a，b，c 可以是单个，也可以是多个。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（Read-Only Memory，简称 ROM）、随机存取存储器（Random Access Memory，简称 RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

权 利 要 求

1、一种动画形象驱动方法，所述方法由音视频处理设备执行，所述方法包括：

获取包含说话人的脸部表情和对应语音的媒体数据；

根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；

根据目标文本信息、所述媒体数据和所述第一表情基，确定对应所述目标文本信息的声学特征和目标表情参数；所述声学特征用于标识模拟所述说话人说出所述目标文本信息的声音，所述目标表情参数用于标识模拟所述说话人说出所述目标文本信息的脸部表情相对于所述第一表情基的变化程度；

根据所述声学特征和所述目标表情参数，驱动具有第二表情基的第二动画形象。

2、根据权利要求 1 所述的方法，所述第一动画形象和所述第二动画形象为同一个动画形象，所述第一表情基与所述第二表情基相同，所述根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，包括：

根据所述脸部表情确定所述第一动画形象的第一表情基和所述第一动画形象的捏脸参数，所述捏脸参数用于标识所述第一动画形象的脸型相对于所述第一动画形象所对应捏脸基的变化程度；

所述根据所述声学特征和所述目标表情参数，驱动具有第二表情基的第二动画形象，包括：

根据所述声学特征、所述目标表情参数和所述捏脸参数，驱动所述第二动画形象。

3、根据权利要求 1 所述的方法，所述第一动画形象和所述第二动画形象为不同动画形象，所述第一表情基与所述第二表情基不同，所述根据所述声学特征和所述目标表情参数，驱动具有第二表情基的第二动画形象，包括：

确定所述第一表情基所对应表情参数与所述第二表情基所对应表情参数间的映射关系；

根据所述声学特征、所述目标表情参数和所述映射关系，驱动所述第二动画形象。

4、根据权利要求 3 所述的方法，所述第二表情基是根据所述第二表情基与音素的预设关系生成的，所述确定所述第一表情基所对应表情参数与所述第二表情基所对应表情参数间的映射关系，包括：

根据所述媒体数据确定所述语音所标识音素、所述音素对应的时间区间和所述媒体数据处于所述时间区间的视频帧；

根据所述视频帧确定所述音素对应的第一表情参数，所述第一表情参数用于标识发出所述音素时所述说话人的脸部表情相对于所述第一表情基的变化程度；

根据所述预设关系和所述第二表情基，确定所述音素对应的第二表情参数；

根据所述第一表情参数和所述第二表情参数，确定所述映射关系。

5、根据权利要求 1 所述的方法，所述根据目标文本信息、所述媒体数据和所述第一表情基，确定对应所述目标文本信息的声学特征和目标表情参数，包括：

根据所述目标文本信息和所述媒体数据，确定对应所述目标文本信息的声学特征和表情特征；所述表情特征用于标识模拟所述说话人说出所述目标文本信息的脸部表情；

根据所述第一表情基和所述表情特征确定所述目标表情参数。

6、一种动画形象驱动装置，所述装置部署在音视频处理设备上，所述装置包括获取单元、第一确定单元、第二确定单元和驱动单元：

所述获取单元，用于获取包含说话人的脸部表情和对应语音的媒体数据；

所述第一确定单元，用于根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；

所述第二确定单元，用于根据目标文本信息、所述媒体数据和所述第一表情基，确定对应所述目标文本信息的声学特征和目标表情参数；所述声学特征用于标识模拟所述说话人说出所述目标文本信息的声音，所述目标表情参数用于标识模拟所述说话人说出所述目标文本信息的脸部表情相对于所述第一表情基的变化程度；

所述驱动单元，用于根据所述声学特征和所述目标表情参数，驱动具有第二表情基的第二动画形象。

7、根据权利要求 6 所述的装置，所述第一动画形象和所述第二动画形象为同一个动画形象，所述第一表情基与所述第二表情基相同，所述第一确定单元，用于：

根据所述脸部表情确定所述第一动画形象的第一表情基和所述第一动画形象的捏脸参数，所述捏脸参数用于标识所述第一动画形象的脸型相对于所述第一动画形象所对应捏脸基的变化程度；

所述驱动单元，用于：

根据所述声学特征、所述目标表情参数和所述捏脸参数，驱动所述第二动画形象。

8、根据权利要求 6 所述的装置，所述第一动画形象和所述第二动画形象为不同动画形象，所述第一表情基与所述第二表情基不同，所述驱动单元，用于：

确定所述第一表情基所对应表情参数与所述第二表情基所对应表情参数间的映射关系；

根据所述声学特征、所述目标表情参数和所述映射关系，驱动所述第二动画形象。

9、根据权利要求 8 所述的装置，所述第二表情基是根据所述第二表情基与音素的预设关系生成的，所述驱动单元，还用于：

根据所述媒体数据确定所述语音所标识音素、所述音素对应的时间区间和所述媒体数据处于所述时间区间的视频帧；

根据所述视频帧确定所述音素对应的第一表情参数，所述第一表情参数用于标识发出所述音素时所述说话人的脸部表情相对于所述第一表情基的变化程度；

根据所述预设关系和所述第二表情基，确定所述音素对应的第二表情参数；

根据所述第一表情参数和所述第二表情参数，确定所述映射关系。

10、根据权利要求 6 所述的装置，所述第二确定单元，用于：

根据所述目标文本信息和所述媒体数据，确定对应所述目标文本信息的声学特征和表情特征；所述表情特征用于标识模拟所述说话人说出所述目标文本信息的脸部表情；

根据所述第一表情基和所述表情特征确定所述目标表情参数。

11、一种动画形象驱动方法，所述方法由音视频处理设备执行，所述方法包括：

获取包含说话人的脸部表情和对应语音的第一媒体数据；

根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；所述第一表情基的维数为第一维数，顶点拓扑为第一顶点拓扑；

根据所述第一表情基和待驱动的第二动画形象的第二表情基，确定目标表情基；所述第二表情基的维数为第二维数，顶点拓扑为第二顶点拓扑，所述目标表情基为具有第二顶点拓扑的第一动画形象对应的表情基，所述目标表情基的维数为第二维数；

根据包含所述说话人的脸部表情和对应语音的第二媒体数据和所述目标表情基，确定目标表情参数和声学特征；所述目标表情参数用于标识所述说话人说出所述语音的脸部表情相对于所述目标表情基的变化程度；

根据所述目标表情参数和声学特征，驱动具有所述第二表情基的所述第二动画形象。

12、根据权利要求 11 所述的方法，所述根据所述第一表情基和待驱动的第二动画形象的第二表情基，确定目标表情基，包括：

从所述第一表情基中确定所述第一动画形象处于无表情时对应的无表情网格，并从所述第二表情基中确所述第二动画形象处于无表情时对应的无表情网格；

根据所述第一动画形象对应的无表情网格和所述第二动画形象对应的无表情网格，确定调整网格，所述调整网格具有第二顶点拓扑，用于标识处于无表情时的第一动画形象；

根据所述调整网格和所述第二表情基中的网格形变关系，生成所述目标表情基。

13、一种动画形象驱动装置，所述装置部署在音视频处理设备上，所述装置包括获取单元、第一确定单元、第二确定单元、第三确定单元和驱动单元：

所述获取单元，用于获取包含说话人的脸部表情和对应语音的第一媒体数据；

所述第一确定单元，用于根据所述脸部表情确定所述说话人所对应第一动画形象的第一表情基，所述第一表情基用于标识所述第一动画形象的表情；所述第一表情基的维数为第一维数，顶点拓扑为第一顶点拓扑；

所述第二确定单元，用于根据所述第一表情基和待驱动的第二动画形象的第二表情基，确定目标表情基；所述第二表情基的维数为第二维数，顶点拓扑为第二顶点拓扑，所述目标表情基为具有第二顶点拓扑的第一动画形象对应的表情基，所述目标表情基的维数为第二维数；

所述第三确定单元，用于根据包含所述说话人的脸部表情和对应语音的第二媒体数据和所述目标表情基，确定目标表情参数和声学特征；所述目标表情参数用于标识所述说话人说出所述语音的脸部表情相对于所述目标表情基的变化程度；

所述驱动单元，用于根据所述目标表情参数和声学特征，驱动具有所述第二表情基的所述第二动画形象。

14、一种用于动画形象驱动的设备，所述设备包括处理器以及存储器：

所述存储器用于存储程序代码，并将所述程序代码传输给所述处理器；

所述处理器用于根据所述程序代码中的指令执行权利要求 1-5 或 11-12 任一项所述的方法。

15、一种计算机可读存储介质，所述计算机可读存储介质用于存储程序代码，所述程序代码用于执行权利要求 1-5 或 11-12 任一项所述的方法。

16、一种计算机程序产品，当所述计算机程序产品被执行时，用于执行权利要求 1-5 或 11-12 任一项所述的方法。

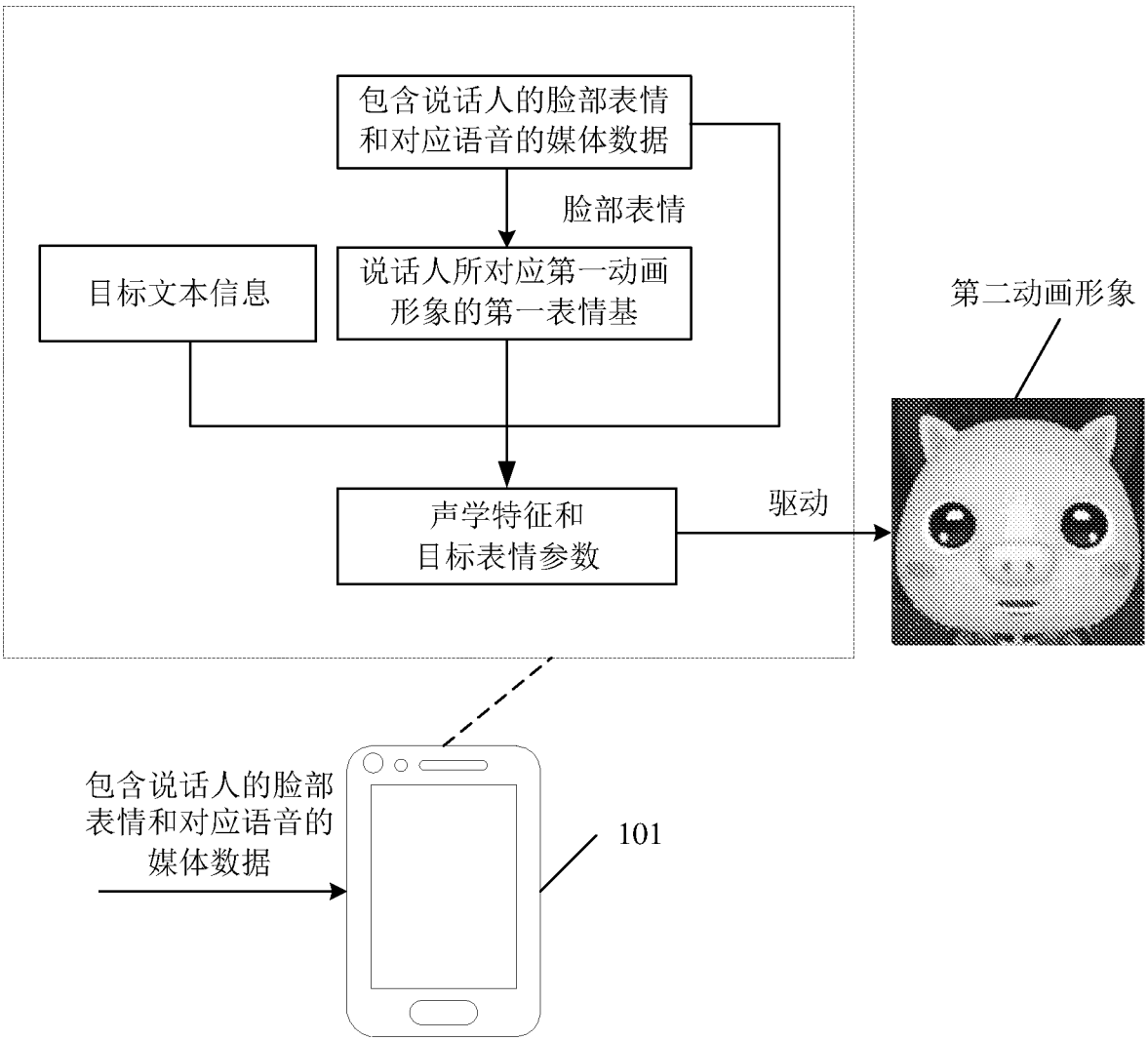

包含说话人的脸部表情和对应语音的媒体数据
脸部表情
目标文本信息
说话人所对应第一动画形象的第一表情基
第二动画形象
声学特征和目标表情参数
驱动
包含说话人的脸部表情和对应语音的媒体数据
101

图 1

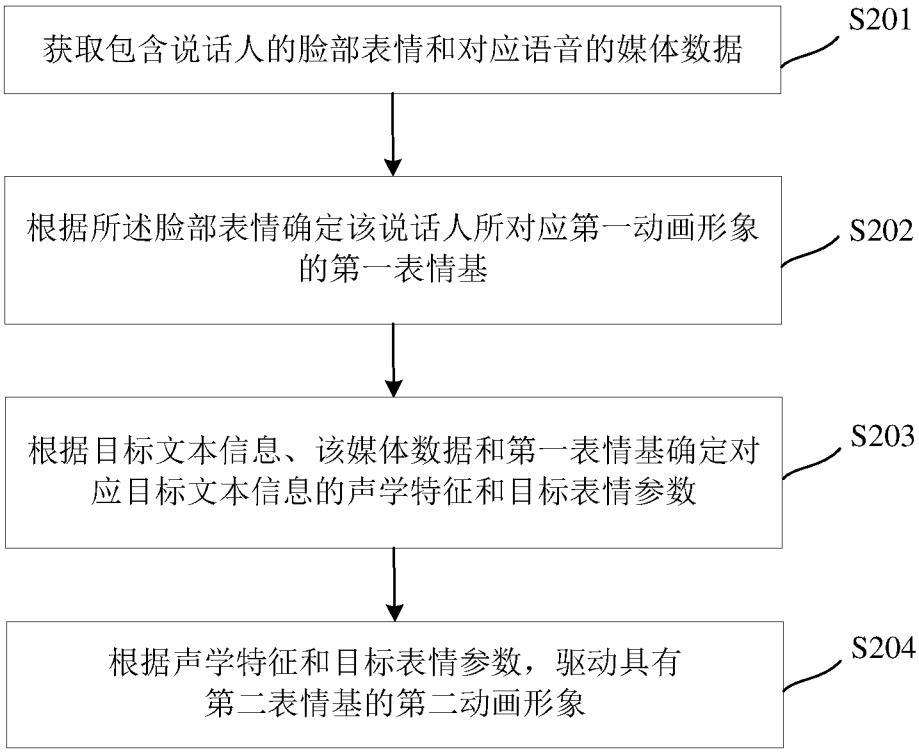

获取包含说话人的脸部表情和对应语音的媒体数据
S201
根据所述脸部表情确定该说话人所对应第一动画形象的第一表情基
S202
根据目标文本信息、该媒体数据和第一表情基确定对应目标文本信息的声学特征和目标表情参数
S203
根据声学特征和目标表情参数，驱动具有第二表情基的第二动画形象
S204

图 2

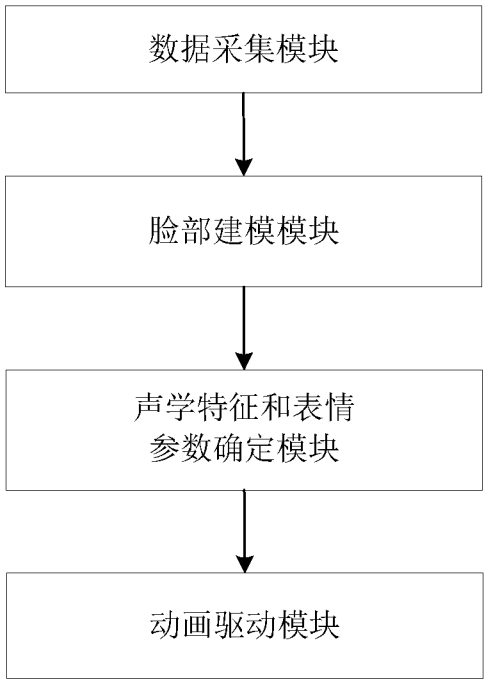

数据采集模块
脸部建模模块
声学特征和表情参数确定模块
动画驱动模块

图 3

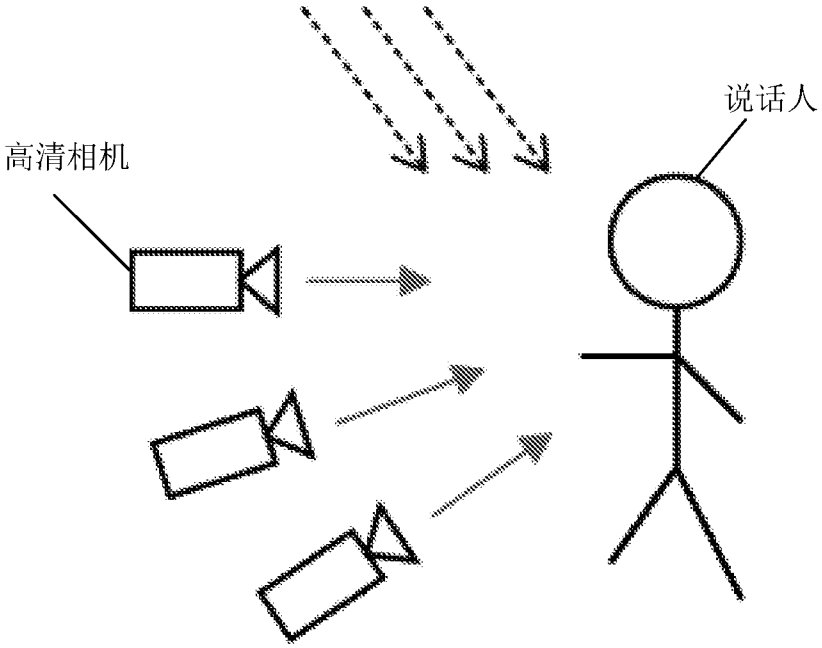


图 4

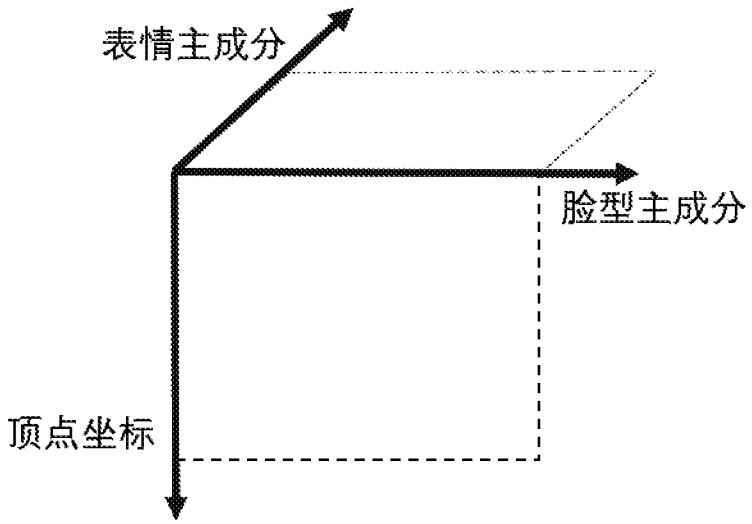


图 5

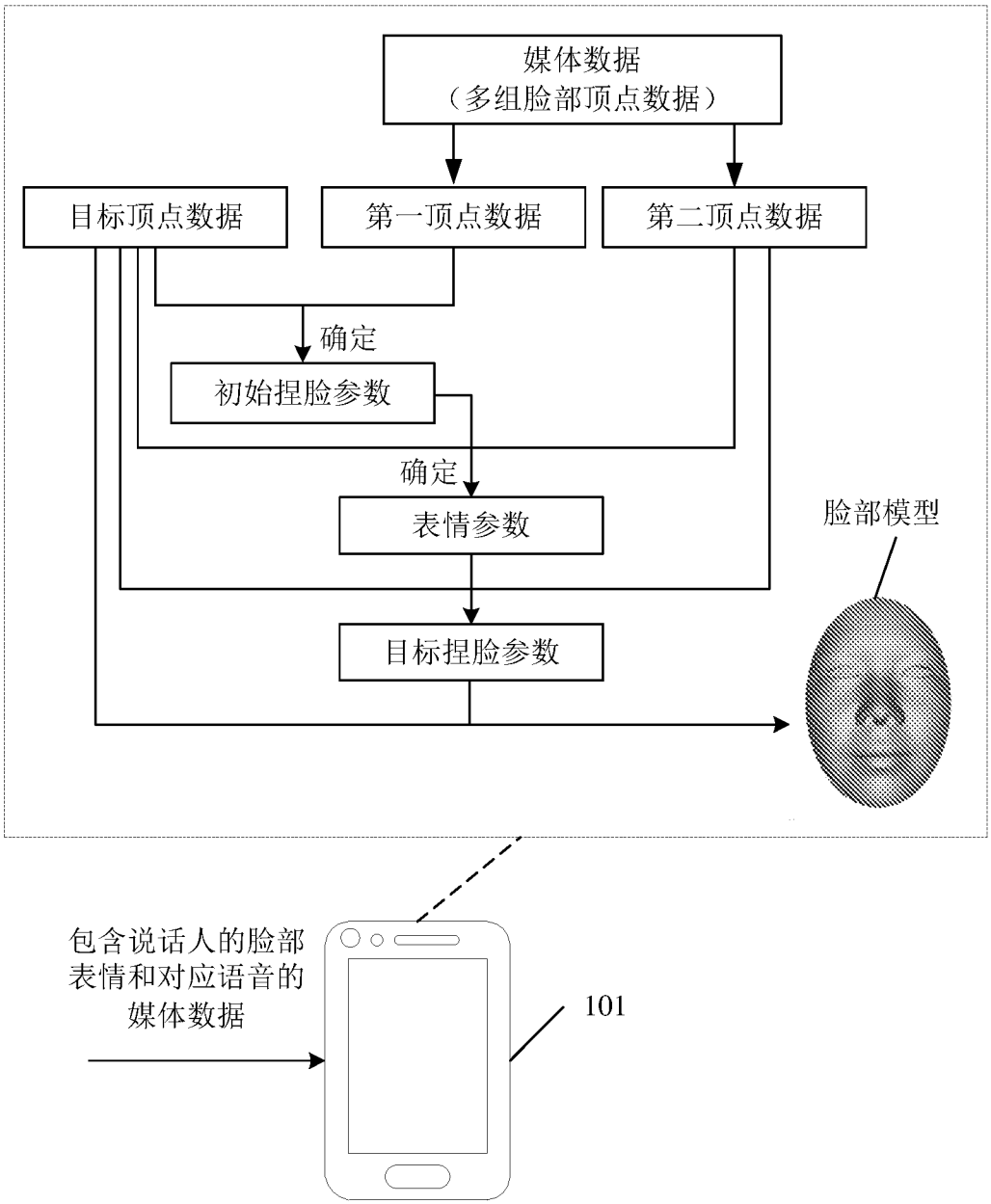

媒体数据
（多组脸部顶点数据）
目标顶点数据
第一顶点数据
第二顶点数据
确定
初始捏脸参数
确定
表情参数
脸部模型
目标捏脸参数
包含说话人的脸部表情和对应语音的媒体数据
101

图 6

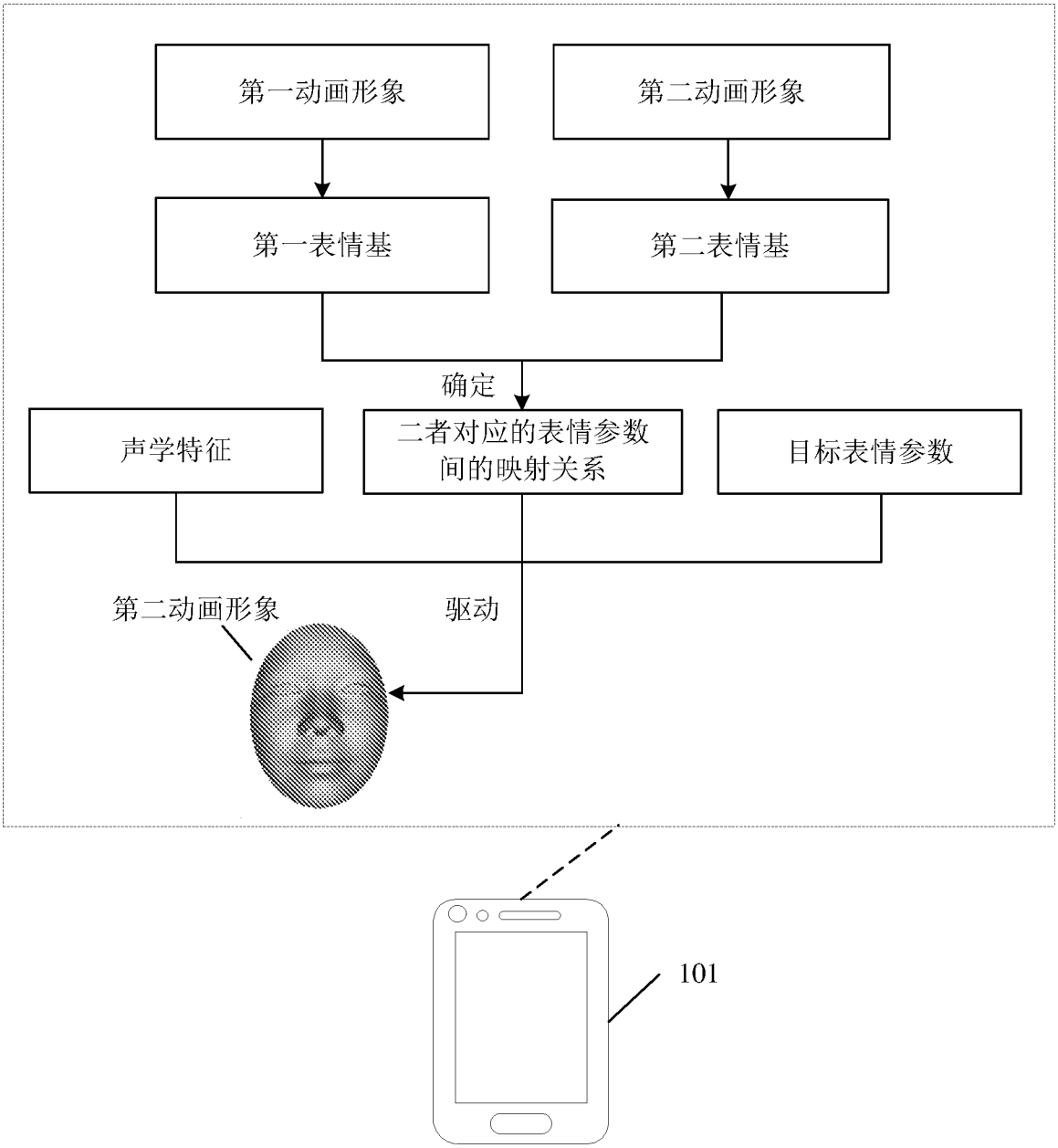

第一动画形象
第二动画形象
第一表情基
第二表情基
确定
声学特征
二者对应的表情参数间的映射关系
目标表情参数
第二动画形象
驱动
101

图 7

```
1   0 5650000 sil
2   5650000 6300000 u
3   6300000 6550000 o
4   6550000 6850000 j
5   6850000 7200000 i
6   7200000 7500000 o
7   7500000 7750000 u
8   7750000 8450000 k
9   8450000 9150000 e
10  9150000 10650000 i
11  10650000 11700000 b
12  11700000 12150000 e
13  12150000 13150000 i
14  13150000 13600000 d
15  13600000 14250000 i
16  14250000 14550000 l
```

图 8

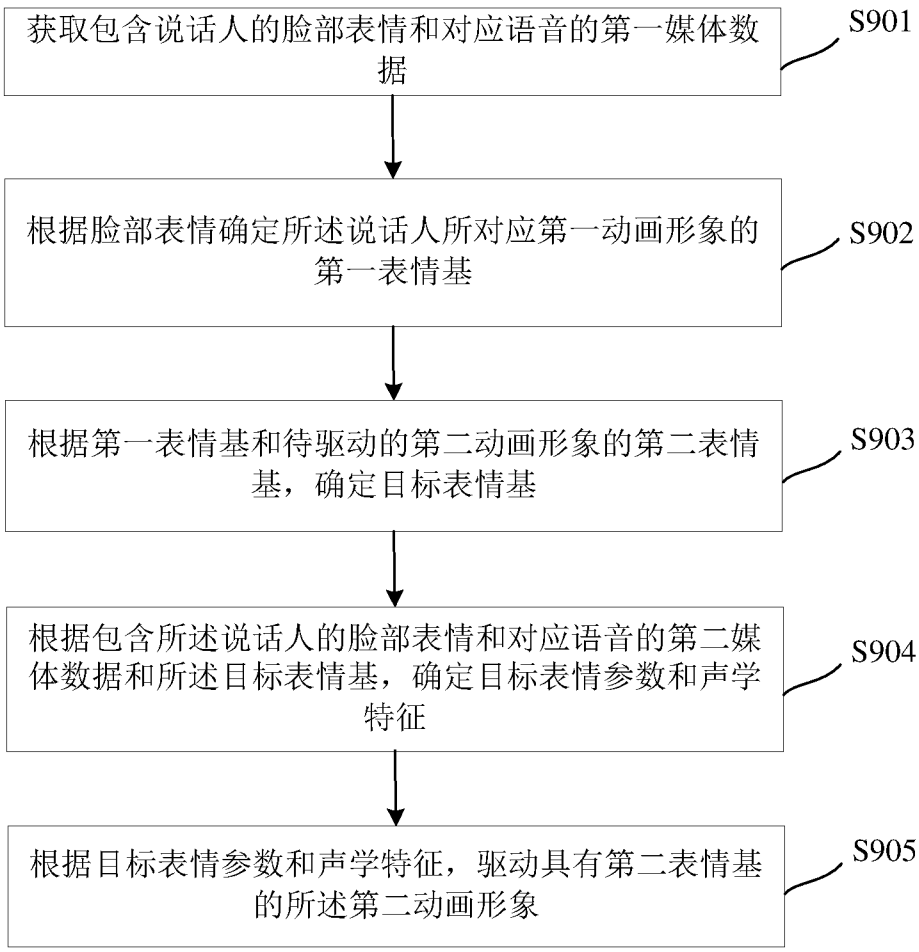


图 9

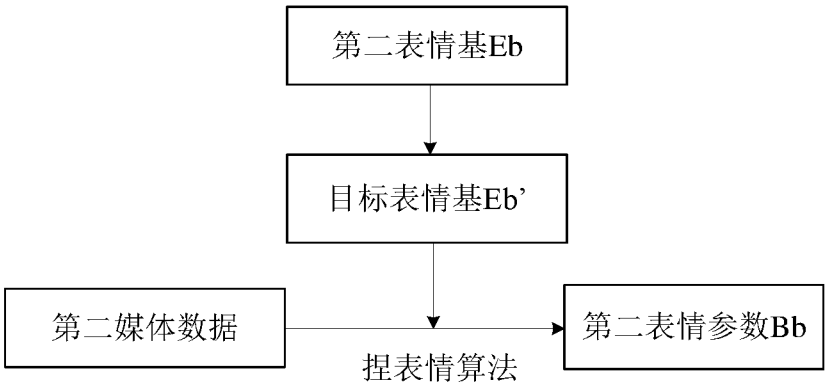


图 10a

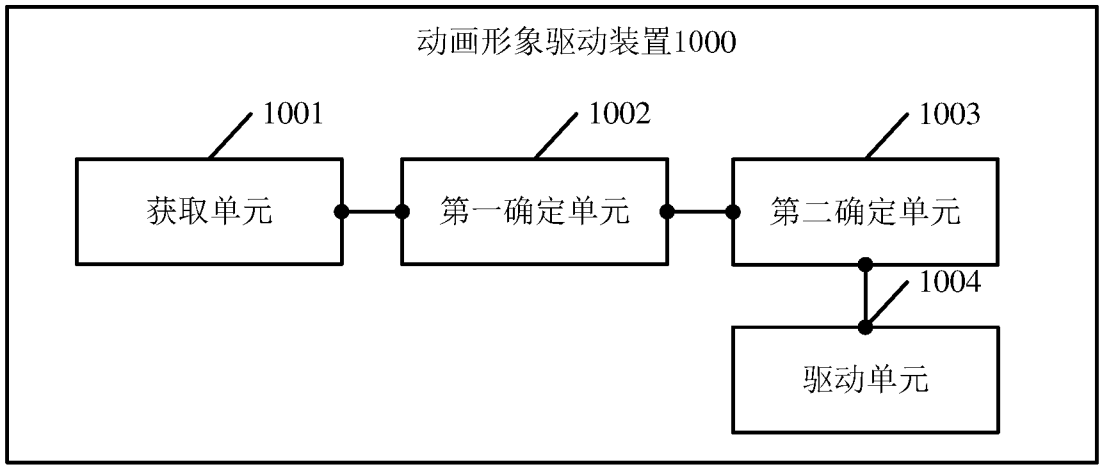


图 10b

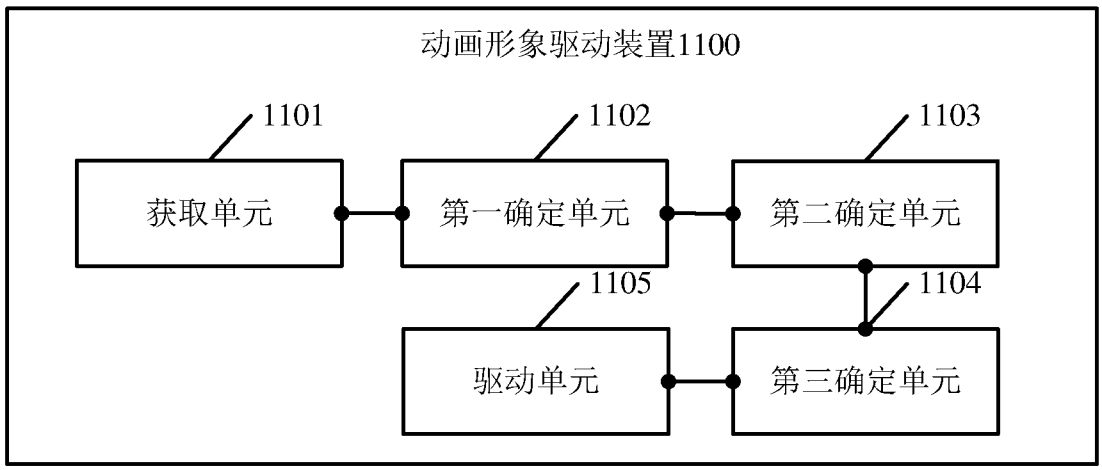


图 11

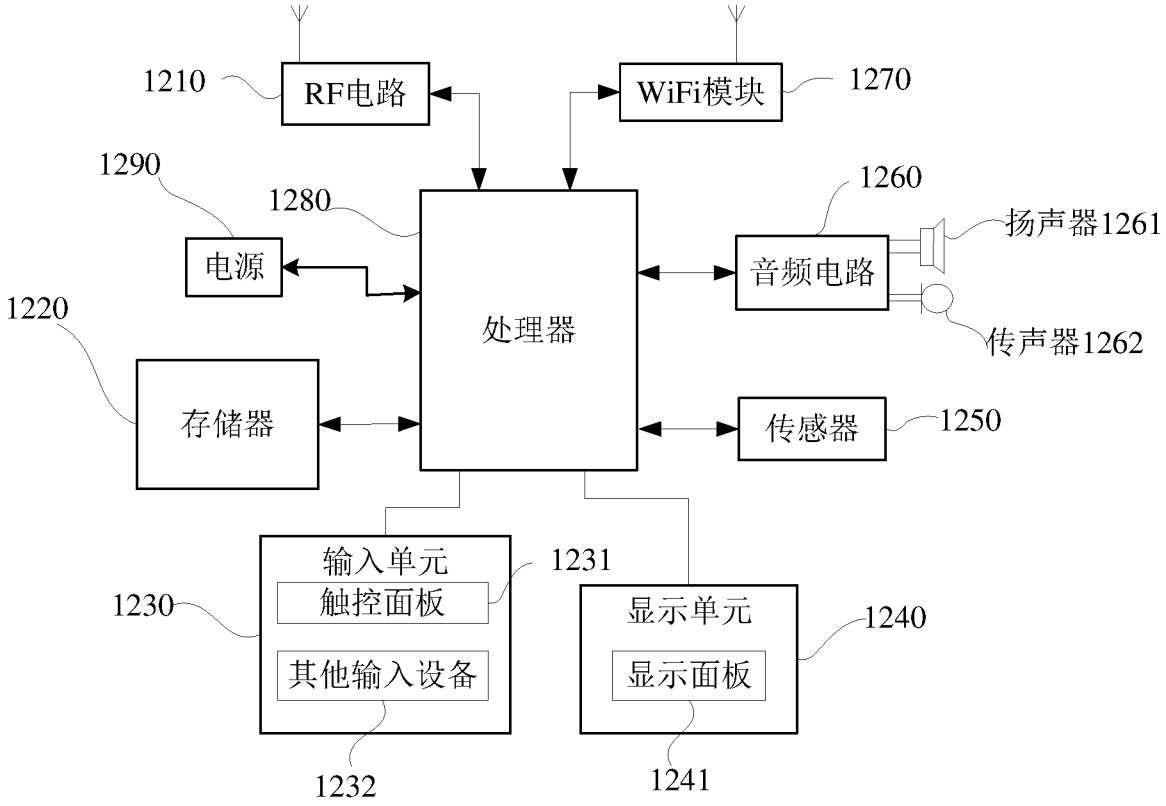

1210
RF电路
WiFi模块
1270
1290
1280
1260
扬声器1261
电源
处理器
音频电路
1220
传声器1262
存储器
传感器
1250
输入单元
1231
1230
触控面板
显示单元
1240
其他输入设备
显示面板
1232
1241

图 12

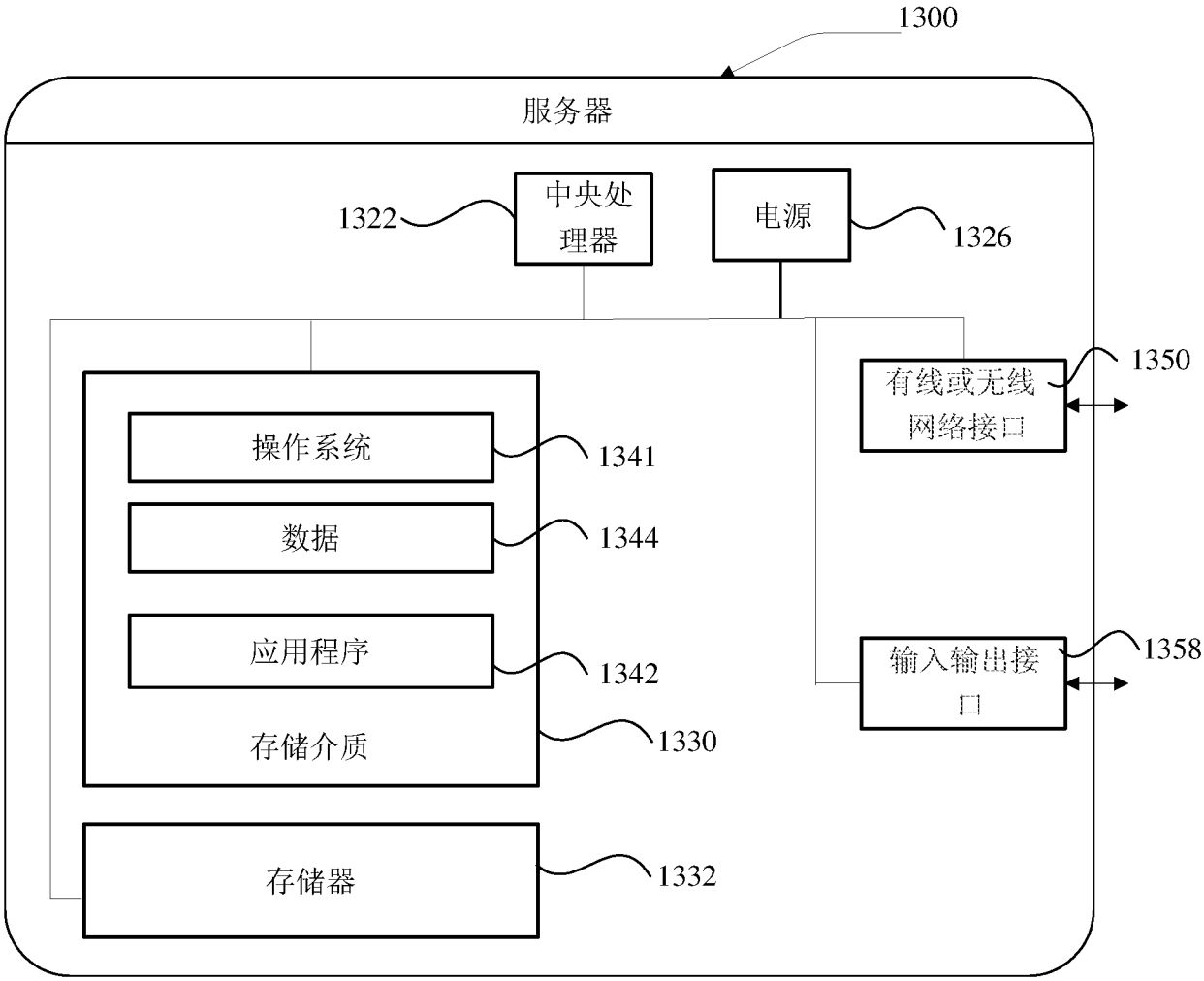

1300
服务器
1322
中央处理器
电源
1326
1350
有线或无线网络接口
操作系统
1341
数据
1344
应用程序
1342
存储介质
1330
输入输出接口
1358
存储器
1332

图 13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/111615

A. CLASSIFICATION OF SUBJECT MATTER

G06F 3/01(2006.01)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)

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)

CNPAT, WPI, EPODOC, CNKI, GOOGLE: 动画, 虚拟, 形象, 驱动, 脸, 表情, 说话, 语音, 视频, animation, visual, driv+, facial, face, voice, expression, video, audio

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110531860 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 03 December 2019 (2019-12-03) claims 1-15	1-16
Y	CN 109447234 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 08 March 2019 (2019-03-08) description, paragraphs [0067]-[0194]	1-3, 5-8, 10, 14-16
Y	CN 104217454 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 17 December 2014 (2014-12-17) description, paragraphs [0044]-[0080]	1-3, 5-8, 10, 14-16
Y	CN 108875633 A (BEIJING KUANGSHI TECHNOLOGY CO., LTD.) 23 November 2018 (2018-11-23) claims 1 and 2	2, 7, 14-16
A	US 8725507 B2 (SAMSUNG ELECTRONICA DA AMAZONIA LTDA.) 13 May 2014 (2014-05-13) entire document	1-16

☑ 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
"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
16 November 2020	**01 December 2020**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/111615

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CAO, Chen et al. "FaceWarehouse: A 3D Facial Expression Database for Visual Computing" *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, Vol. 20, No. 3, 31 March 2014 (2014-03-31), entire document	1-16

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2020/111615

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110531860	A	03 December 2019	CN	110531860	B	24 July 2020
CN	109447234	A	08 March 2019	CN	110288077	A	27 September 2019
CN	104217454	A	17 December 2014	CN	104217454	B	03 November 2017
CN	108875633	A	23 November 2018	US	2019384967	A1	19 December 2019
US	8725507	B2	13 May 2014	BR	PI0904540	A2	12 July 2011
				US	2011131041	A1	02 June 2011

国际检索报告

国际申请号
PCT/CN2020/111615

A. 主题的分类

G06F 3/01(2006.01)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)

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））

CNPAT, WPI, EPODOC, CNKI, GOOGLE: 动画，虚拟，形象，驱动，脸，表情，说话，语音，视频，animation, visual, driv+, facial, face, voice, expression, video, audio

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 110531860 A (腾讯科技深圳有限公司) 2019年 12月 3日 (2019-12-03) 权利要求1-15	1-16
Y	CN 109447234 A (腾讯科技深圳有限公司) 2019年 3月 8日 (2019-03-08) 说明书第[0067]-[0194]段	1-3，5-8，10，14-16
Y	CN 104217454 A (中国科学院计算技术研究所) 2014年 12月 17日 (2014-12-17) 说明书第[0044]-[0080]段	1-3，5-8，10，14-16
Y	CN 108875633 A (北京旷视科技有限公司) 2018年 11月 23日 (2018-11-23) 权利要求1-2	2，7，14-16
A	US 8725507 B2 (SAMSUNG ELECTRONICA DA AMAZONIA LTDA.) 2014年 5月 13日 (2014-05-13) 全文	1-16
A	CAO, Chen 等. "FaceWarehouse: A 3D Facial Expression Database for Visual Computing" IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 第20卷，第3期，2014年 3月 31日 (2014-03-31)， 全文	1-16

☐ 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

“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
2020年 11月 16日	2020年 12月 1日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 胡丽丽 电话号码 86-(10)-53961386

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/111615

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	110531860	A	2019年 12月 3日	CN	110531860	B	2020年 7月 24日
CN	109447234	A	2019年 3月 8日	CN	110288077	A	2019年 9月 27日
CN	104217454	A	2014年 12月 17日	CN	104217454	B	2017年 11月 3日
CN	108875633	A	2018年 11月 23日	US	2019384967	A1	2019年 12月 19日
US	8725507	B2	2014年 5月 13日	BR	PI0904540	A2	2011年 7月 12日
				US	2011131041	A1	2011年 6月 2日