(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110503942 A
(43)申请公布日 2019.11.26

(21)申请号 201910820742.1

(22)申请日 2019.08.29

(71)申请人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 康世胤 陀得意 李广之 傅天晓 黄晖榕 苏丹

(74)专利代理机构 深圳市深佳知识产权代理事务所(普通合伙) 44285
代理人 王兆林

(51)Int.Cl.
G10L 15/02(2006.01)
G10L 15/22(2006.01)
G10L 15/25(2013.01)

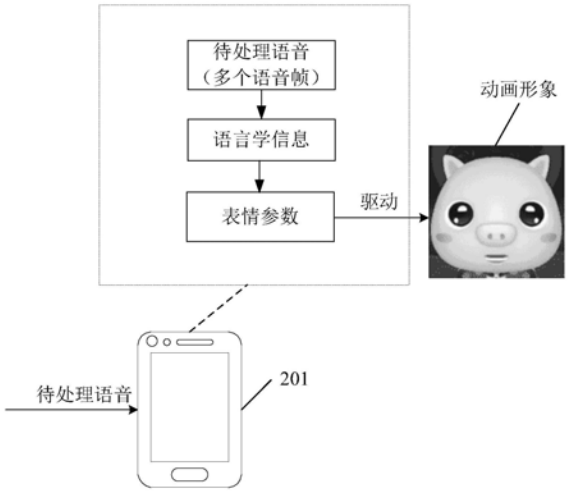
权利要求书2页 说明书13页 附图6页

(54)发明名称

一种基于人工智能的语音驱动动画方法和装置

(57)摘要

本申请实施例公开了一种基于人工智能的语音驱动动画方法,当获取包括多个语音帧的待处理语音,可以确定出待处理语音中语音帧对应的语言学信息,每一个语言学信息用于标识所对应语音帧所属音素的分布可能性,即体现语音帧中内容属于哪一种音素的概率分布,该语言学信息携带的信息与待处理语音的实际说话人无关,由此可以抵消不同说话人发音习惯对后续表情参数的确定所带来的影响,根据语言学信息所确定出的表情参数,可以准确驱动动画形象做出对应待处理语音的表情,例如口型,从而可以有效支持任意说话人对应的待处理语音,提高了交互体验。



CN 110503942 A

1.一种语音驱动动画方法,其特征在于,所述方法包括:

获取待处理语音,所述待处理语音包括多个语音帧;

确定所述待处理语音中语音帧对应的语言学信息,所述语言学信息用于标识所述待处理语音中语音帧所属音素的分布可能性;

根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数;

根据所述表情参数驱动动画形象做出对应所述待处理语音的表情。

2.根据权利要求1所述的方法,其特征在于,目标语音帧为所述待处理语音中的一个语音帧,针对所述目标语音帧,所述根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数,包括:

确定所述目标语音帧所处的语音帧集合,所述语音帧集合包括所述目标语音帧和多个语音帧,所述多个语音帧为所述目标语音帧的上下文语音帧;

根据所述语音帧集合中语音帧分别对应的语言学信息,确定所述目标语音帧对应的表情参数。

3.根据权利要求2所述的方法,其特征在于,所述语音帧集合中语音帧的数量是根据神经网络映射模型确定的,或者,所述语音帧集合中语音帧的数量是根据所述待处理语音的语音切分结果确定。

4.根据权利要求2所述的方法,其特征在于,所述多个语音帧为所述目标语音帧的相邻上下文语音帧,或者,所述多个语音帧为所述目标语音帧的间隔上下文语音帧。

5.根据权利要求2所述的方法,其特征在于,所述根据所述语音帧集合中语音帧分别对应的语言学信息确定所述目标语音帧对应的表情参数,包括:

根据所述语音帧集合中语音帧分别对应的语言学信息,确定所述语音帧集合中语音帧分别对应的待定表情参数;

根据所述目标语音帧在不同语音帧集合中分别确定的待定表情参数,计算所述目标语音帧对应的表情参数。

6.根据权利要求1-5任意一项所述的方法,其特征在于,所述语言学信息包括音素后验概率、瓶颈特征和嵌入特征中任意一种或多种的组合。

7.根据权利要求1-5任意一项所述的方法,其特征在于,所述根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数,包括:

根据所述语言学信息,通过神经网络映射模型确定所述待处理语音中语音帧对应的表情参数;所述神经网络映射模型包括深度神经网络DNN模型、长短期记忆网络LSTM模型或双向长短期记忆网络BLSTM模型。

8.根据权利要求1-5任意一项所述的方法,其特征在于,所述根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数,包括:

根据所述语言学信息和对应的情感向量,确定所述待处理语音中语音帧对应的表情参数。

9.根据权利要求1-5任意一项所述的方法,其特征在于,所述确定所述待处理语音中语音帧对应的语言学信息,包括:

确定所述待处理语音中语音帧对应的声学特征;

通过自动语音识别模型确定所述声学特征对应的语言学信息。

10.根据权利要求9所述的方法，其特征在于，所述自动语音识别模型是根据包括了语音片段和音素对应关系的训练样本训练得到的。

11.一种语音驱动动画装置，其特征在于，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，用于获取待处理语音，所述待处理语音包括多个语音帧；

所述第一确定单元，用于确定所述待处理语音中语音帧对应的语言学信息，所述语言学信息用于标识所述待处理语音中语音帧所属音素的分布可能性；

所述第二确定单元，用于根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数；

所述驱动单元，用于根据所述表情参数驱动动画形象做出对应所述待处理语音的表情。

12.根据权利要求11所述的装置，其特征在于，目标语音帧为所述待处理语音中的一个语音帧，针对所述目标语音帧，所述第二确定单元，用于：

确定所述目标语音帧所处的语音帧集合，所述语音帧集合包括所述目标语音帧和多个语音帧，所述多个语音帧为所述目标语音帧的上下文语音帧；

根据所述语音帧集合中语音帧分别对应的语言学信息，确定所述目标语音帧对应的表情参数。

13.根据权利要求10所述的装置，其特征在于，所述语音帧集合中语音帧的数量是根据神经网络映射模型确定的，或者，所述语音帧集合中语音帧的数量是根据所述待处理语音的语音切分结果确定。

14.一种设备，其特征在于，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1-10任意一项所述的方法。

15.一种计算机可读存储介质，其特征在于，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1-10任意一项所述的方法。

一种基于人工智能的语音驱动动画方法和装置

技术领域

[0001] 本申请涉及数据处理领域，特别是涉及一种基于人工智能的语音驱动动画方法和装置。

背景技术

[0002] 目前，语音到虚拟人脸动画的生成这一技术正在成为工业界应用领域的研究热点，例如针对一段任意说话人的语音，可以驱动一个动画形象做出该段语音对应的口型。在这一场景下，动画形象的存在能极大地增强真实感，提升表现力，带给用户更加沉浸式的体验。

[0003] 一种方式是通过Speech2Face系统实现上述技术。一般来说，针对一说话人的语音，该系统提取语音中的声学特征例如梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient，MFCC)后，通过映射模型可以基于声学特征确定出对应于一个能够调整的动画形象的表情参数，依据该表情参数可以控制该动画形象做出该段语音对应的口型。

[0004] 然而，由于提取的声学特征中含有与说话人相关的信息，导致以此建立的映射模型对特定说话人的语音可以准确确定对应的表情参数，若说话人出现了变更，映射模型确定出的表情参数将出现较大偏差，以此驱动的动画形象口型将与语音不一致，降低了交互体验。

发明内容

[0005] 为了解决上述技术问题，本申请提供了基于人工智能的语音驱动动画方法和装置，可以有效支持任意说话人对应的待处理语音，提高了交互体验。

[0006] 本申请实施例公开了如下技术方案：

[0007] 第一方面，本申请实施例提供一种语音驱动动画方法，所述方法包括：

[0008] 获取待处理语音，所述待处理语音包括多个语音帧；

[0009] 确定所述待处理语音中语音帧对应的语言学信息，所述语言学信息用于标识所述待处理语音中语音帧所属音素的分布可能性；

[0010] 根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数；

[0011] 根据所述表情参数驱动动画形象做出对应所述待处理语音的表情。

[0012] 第二方面，本申请实施例提供一种语音驱动动画装置，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：

[0013] 所述获取单元，用于获取待处理语音，所述待处理语音包括多个语音帧；

[0014] 所述第一确定单元，用于确定所述待处理语音中语音帧对应的语言学信息，所述语言学信息用于标识所述待处理语音中语音帧所属音素的分布可能性；

[0015] 所述第二确定单元，用于根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数；

[0016] 所述驱动单元，用于根据所述表情参数驱动动画形象做出对应所述待处理语音的

表情。

[0017] 第三方面，本申请实施例提供一种设备，所述设备包括处理器以及存储器：

[0018] 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；

[0019] 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所述的方法。

[0020] 第四方面，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所述的方法。

[0021] 由上述技术方案可以看出，当获取包括多个语音帧的待处理语音，可以确定出待处理语音中语音帧对应的语言学信息，每一个语言学信息用于标识所对应语音帧所属音素的分布可能性，即体现语音帧中内容属于哪一种音素的概率分布，该语言学信息携带的信息与待处理语音的实际说话人无关，由此可以抵消不同说话人发音习惯对后续表情参数的确定所带来的影响，根据语言学信息所确定出的表情参数，可以准确驱动动画形象做出对应待处理语音的表情，例如口型，从而可以有效支持任意说话人对应的待处理语音，提高了交互体验。

附图说明

[0022] 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。

[0023] 图1为相关技术中所采用的Speech2Face系统；

[0024] 图2为本申请实施例提供的一种语音驱动动画方法的应用场景示意图；

[0025] 图3为本申请实施例提供的一种语音驱动动画方法的流程图；

[0026] 图4为本申请实施例提供的Speech2Face的系统架构；

[0027] 图5为本申请实施例提供的ASR模型的训练过程示例图；

[0028] 图6为本申请实施例提供的基于DNN模型确定表情参数的示例图；

[0029] 图7为本申请实施例提供的相邻上下文语音帧的示例图；

[0030] 图8为本申请实施例提供的间隔上下文语音帧的示例图；

[0031] 图9为本申请实施例提供的基于LSTM模型确定表情参数的示例图；

[0032] 图10为本申请实施例提供的基于BLSTM模型确定表情参数的示例图；

[0033] 图11为本申请实施例提供的一种语音驱动动画装置的结构图；

[0034] 图12为本申请实施例提供的终端设备的结构图；

[0035] 图13为本申请实施例提供的服务器的结构图。

具体实施方式

[0036] 下面结合附图，对本申请的实施例进行描述。

[0037] 相关技术中所采用的Speech2Face系统参见图1所示。针对一说话人的语音，该系统可以对语音进行声学特征提取，得到MFCC。然后，通过映射模型基于声学特征确定出表情参数。对于设定好的、可以通过调整表情参数来调整表情(例如口型)的动画形象，利用确定出的表情参数调整动画形象，生成该段语音对应的动画形象。

[0038] 然而，由于相关技术中提取的声学特征与说话人相关，当说话人更换时，映射模型确定出的表情参数将出现较大偏差，以此驱动的动画形象口型将与语音不一致，降低了交互体验。

[0039] 为此，本申请实施例提供一种基于人工智能的语音驱动动画方法，该方法在获取到包括多个语音帧的待处理语音后，可以确定出待处理语音中语音帧对应的语言学信息。与相关技术中提取的声学特征例如MFCC相比，语言学信息携带的信息与待处理语音的实际说话人无关，避免了不同说话人发音习惯对后续表情参数的确定所带来的影响，故，针对任意说话人对应的待处理语音，可以根据语言学信息确定表情参数，从而准确驱动动画形象做出对应待处理语音的表情。

[0040] 需要强调的是，本申请实施例所提供的语音驱动动画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。

[0041] 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。

[0042] 在本申请实施例中，主要涉及的人工智能软件技术包括上述语音处理技术和机器学习等方向。

[0043] 例如可以涉及语音技术(Speech Technology)中的语音识别技术，其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。

[0044] 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。

[0045] 本申请实施例提供的基于人工智能的语音驱动动画方法可以应用于具有驱动动画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。

[0046] 该音视频处理设备可以具有实施语音技术中自动语音识别技术(ASR)和声纹识别的能力。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。

[0047] 在本申请实施例中，音视频处理设备通过实施上述语音技术，可以对获取的待处理语音进行识别确定待处理语音中语音帧对应的语言学信息等功能；通过机器学习技术训练神经网络映射模型，通过训练得到的神经网络映射模型根据语言学信息确定表情参数，以驱动动画形象做出对应待处理语音的表情。

[0048] 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。

[0049] 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。当服务器实施该语音驱动动画方法时，服务器确定出表情参数，利用该表情参数驱动终端设备上的动画形象做出对应待处理语音的表情。

[0050] 需要说明的是，本申请实施例提供的语音驱动动画方法可以应用到多种承担一些人类工作的应用场景，例如新闻播报、天气预报以及游戏解说等，还能承担一些私人化的服务，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驱动动画形象做出表情，极大地增强真实感，提升表现力。

[0051] 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语音驱动动画方法进行介绍。

[0052] 参见图2，图2为本申请实施例提供的基于人工智能的语音驱动动画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201，终端设备201可以获取待处理语音，待处理语音中包括多个语音帧。待处理语音为与任意说话人对应的语音，例如为说话人发出的语音。本实施例并不限定待处理语音的类型，其中，待处理语音可以是说话人说话所对应的语音，也可以是说话人唱歌所对应的语音。本实施例也不限定待处理语音的语种，例如待处理语音可以是汉语、英语等。

[0053] 可以理解的是，待处理语音不仅可以为说话人通过终端设备201输入的语音。在一些情况下，本申请实施例提供的方法所针对的待处理语音也可以是根据文本生成的语音，即通过终端设备201输入文本，通过智能语音平台转换成符合说话人语音特点的语音，将该语音作为待处理语音。

[0054] 由于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、无论待处理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，只要待处理语音中语音帧对应的音素相同，那么，对应的表情例如口型具有一致性。基于音素的特点，在本实施例中，终端设备201可以确定待处理语音中语音帧对应的语言学信息，语言学信息用于标识待处理语音中语音帧所属音素的分布可能性，即语音帧内容属于哪一种音素的概率分布，从而确定语音帧内容所属音素。

[0055] 可见，与前述相关内容中所涉及的相关技术中的声学特征相比，语言学信息携带的信息与待处理语音的实际说话人无关，不管说话人是谁、语音帧内容所对应的文本是什么，语音帧内容所属的音素(语言学信息)是可以确定的，例如确定出语音帧内容所属的音素是“a”，虽然音素“a”对应的说话人可能不同，对应的文本也有可能不同，但是，只要发出的是音素“a”，音素“a”对应的表情例如口型一致的。故，终端设备201可以根据语言学信息准确地确定出表情参数，从而准确驱动动画形象做出对应待处理语音的表情，避免了不同说话人发音习惯对表情参数的确定所带来的影响，提高了交互体验。

[0056] 接下来,将结合附图对本申请实施例提供的语音驱动动画方法进行详细介绍。

[0057] 参见图3,图3示出了一种语音驱动动画方法的流程图,所述方法包括:

[0058] S301、获取待处理语音。

[0059] 以音视频处理设备是终端设备为例,当说话人通过麦克风向终端设备输入待处理语音,以希望根据待处理语音驱动动画形象做出对应待处理语音的表情时,终端设备可以获取该待处理语音。本申请实施例对说话人、待处理语音类型、语种等不做限定。该方法可以支持任意说话人对应的语音,即任意说话人对应的语音可以作为待处理语音;该方法可以支持多种语种,即待处理语音的语种可以是汉语、英语、法语等各种语种;该方法还可以支持唱歌语音,即待处理语音可以是说话人唱歌的语音。

[0060] S302、确定所述待处理语音中语音帧对应的语言学信息。

[0061] 其中,语言学信息是与说话人无关的信息,用于标识所述待处理语音中语音帧所属音素的分布可能性。在本实施例中,语言学信息可以包括音素后验概率(Phonetic Posterior grams,PPG)、瓶颈(bottomneck)特征和嵌入(imbedding)特征中任意一种或多种的组合。后续实施例中主要以语言学信息是PPG为例进行介绍。

[0062] 需要说明的是,本申请实施例实现语音驱动动画时采用的也是Speech2Face系统,但是本申请实施例采用的Speech2Face系统与前述内容相关技术中的Speech2Face系统有所不同,本申请实施例提供的Speech2Face的系统架构可以参考图4所示。其中,主要由四部分组成,第一部分是训练得到自动语音识别(Automatic Speech Recognition,ASR)模型,用于PPG提取;第二部分是基于训练好的ASR模型,确定待处理语音的PPG(例如S302);第三部分则是声学参数到表情参数的映射,即基于音素后验概率确定待处理语音中语音帧对应的表情参数(例如S303);第四部分是根据表情参数驱动设计好的3D动画形象做出对应待处理语音的表情(例如S304)。

[0063] 终端设备对待处理语音进行语言学信息特征提取,从而确定出待处理语音中语音帧对应的语言学信息。

[0064] 本实施例中所涉及的音素可以包括218种,包括汉语音素、英语音素等多个语种,从而实现多语种的语言学信息提取。若语言学信息是PPG,故,得到的PPG可以是一个218维的向量。

[0065] 在一种实现方式中,终端设备可以通过ASR模型实现语言学信息提取。在这种情况下,以语言学信息是PPG为例,为了能够提取PPG,需要预先训练一个ASR模型(即上述第一部分)。ASR模型的训练方式可以是根据包括了语音片段和音素对应关系的训练样本训练得到的。具体的,ASR模型是基于Kaldi所提供的ASR接口,在给定ASR数据集的情况下进行训练得到,ASR数据集中包括了训练样本。其中,Kaldi是一种开源的语音识别工具箱,Kaldi使用的是一个基于深度置信网络-深度神经网络(Deep Belief Network-Deep neural network,DBN-DNN)的网络结构来根据提取的MFCC预测语音片段属于每个音素的可能性大小,也即PPG,从而对输出的音素进行分类,ASR模型的训练过程可以参见图5中虚线所示。当训练得到上述ASR模型之后,对于待处理语音,经过上述ASR模型之后,就能够输出得到待处理语音中语音帧对应的PPG,从而可以用于后续表情参数的确定。其中,ASR模型实际上是DNN模型。

[0066] 基于ASR模型的训练方式,利用ASR模型确定语言学信息的方式可以是确定待处理语音中语音帧对应的声学特征,该声学特征为前述相关内容中所涉及的相关技术中的声学

特征，例如MFCC。然后利用ASR模型确定该声学特征对应的语言学信息。

[0067] 需要说明的是，由于训练ASR模型所使用的ASR数据集就考虑了带噪声的语音片段情况，对噪声的适应性更强，因此通过ASR模型提取的语言学信息相比MFCC等相关技术中所使用的声学特征来说鲁棒性更强。

[0068] S303、根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数。表情参数用于驱动动画形象做出对应待处理语音的表情，即通过确定出的表情参数对预先建立的动画形象的表情参数进行调整，从而使得动画形象做出与说出待处理语音相符的表情。

[0069] 应理解，对于动画形象而言，表情可以包括面部表情和身体姿态表情，面部表情例如可以包括口型、五官动作和头部姿态等，身体姿态表情例如可以包括身体动作、手势动作以及走路姿态等等。

[0070] 在一种实现方式中，S303的实现方式可以是根据语言学信息，通过神经网络映射模型确定待处理语音中语音帧对应的表情参数，例如图4中虚线框所示。其中，神经网络映射模型可以包括DNN模型、长短期记忆网络(Long Short-Term Memory，LSTM)模型或双向长短期记忆网络(Bidirectional Long Short-term Memory，BLSTM)模型。

[0071] 神经网络映射模型是预先训练得到的，神经网络映射模型实现的是语言学信息到表情参数的映射，即当输入语言学信息，便可以输出待处理语音中语音帧对应的表情参数。

[0072] 由于语言学信息与表情中的口型较为相关，通过语言学信息确定口型更为精确。而口型之外的其他表情与待处理语音对应的情感更为相关，为了准确的确定表情参数，以便通过表情参数驱动动画形象做出更为丰富的表情，例如，在做出口型的同时还可以大笑、眨眼等。S303的一种可能实现方式可以参见图5所示(图5以语言学信息是PPG为例)，在利用训练好的ASR模型得到语音帧的PPG之后，将PPG与预先标注好的情感向量进行拼接得到最终的特征，从而根据PPG和对应的情感向量，确定待处理语音中语音帧对应的表情参数。

[0073] 其中，在情感方面，本实施例中采用了四种常用的情感，包括快乐，悲伤，生气和正常状态。利用情感向量来表征情感，情感向量采用了1-of-K编码方式，即设定长度为4，在四个维度上分别取1，其他维度取0得到四个向量，用以分别表示四种情感。在确定出语音帧的PPG后，即得到一个218维的向量，与该待处理语音的4维情感向量进行拼接就能得到一个222维的特征向量，用于后续作为神经网络映射模型的输入。

[0074] 需要说明的是，在本实施例中所使用的基于神经网络的神经网络映射模型，还可以使用Tacotron decoder进行替换。其中，Tacotron decoder是一种注意力模型，用于端到端的语音合成。

[0075] S304、根据所述表情参数驱动动画形象做出对应所述待处理语音的表情。

[0076] 对于动画形象本身而言，动画形象可以是3D的形象也可以是2D的形象，本实施例对此不做限定。例如，建立的动画形象如图2中的动画形象所示，利用该动画形象向大家拜年，假设当待处理语音为“恭喜发财，大吉大利”，那么，根据该待处理语音可以确定出表情参数，从而驱动该动画形象做出可以发出“恭喜发财，大吉大利”的表情(口型)。当然，还可以做出除了口型之外的其他表情。

[0077] 由上述技术方案可以看出，当获取包括多个语音帧的待处理语音，可以确定出待处理语音中语音帧对应的语言学信息，每一个语言学信息用于标识所对应语音帧所属音素的分布可能性，即体现语音帧中内容属于哪一种音素的概率分布，该语言学信息携带的信

息与待处理语音的实际说话人无关，由此可以抵消不同说话人发音习惯对后续表情参数的确定所带来的影响，根据语言学信息所确定出的表情参数，可以准确驱动动画形象做出对应待处理语音的表情，例如口型，从而可以有效支持任意说话人对应的待处理语音，提高了交互体验。

[0078] 待处理语音中包括多个语音帧，接下来将以待处理语音中的一个语音帧作为目标语音帧为例，详细介绍针对目标语音帧，S303如何确定出目标语音帧对应的表情参数。

[0079] 由于语音存在协同发音的效应，目标语音帧对应的口型等表情会与其上下文语音帧短时相关，因此，为了可以准确的确定出目标语音帧对应的表情参数可以在确定表情参数时结合上下文语音帧，即确定目标语音帧所处的语音帧集合，该语音帧集合中包括了目标语音帧和多个语音帧，该多个语音帧为目标语音帧的上下文语音帧，从而根据语音帧集合中语音帧分别对应的语言学信息确定目标语音帧对应的表情参数。

[0080] 可以理解的是，若使用神经网络映射模型确定表情参数，根据所使用的神经网络映射模型的不同，确定语音帧集合的方式以及语音帧集合中语音帧的数量有所不同。若神经网络映射模型是DNN模型，即基于前向连接的分类器，如图6所示，由于DNN模型的输入要求是定长，因此考虑逐帧输入进行预测而不是输入变长的序列去做序列预测。由于需要结合上下文语音帧的信息来确定目标语音帧的表情参数，为了达到引入上下文语音帧的目的，输入不再是一帧，而是对输入的待处理语音以目标语音帧为中心加窗，将窗中的语音帧一起作为短时序列输入，此时，窗中的语音帧包括目标语音帧和多个语音帧，构成语音帧集合。可见，通过选取固定的窗长，就能够自然地满足DNN模型的输入要求。

[0081] 在这种情况下，语音帧集合中语音帧的数量是由窗长决定的，而窗长反映神经网络映射模型的输入要求，即语音帧集合中语音帧的数量是根据神经网络映射模型确定的。

[0082] 以图6为例，若PPG通过218维向量表示，窗长为7，则输入DNN模型的输入为包括目标语音帧的7个语音帧，即语音帧集合中语音帧的数量为7帧，每一个语音帧对应一个PPG(共218维向量)，7帧则对应218×7维向量，图6中每个圆圈表示一维参数。当输入7个语音帧分别对应的PPG时，便可输出目标语音帧对应的表情参数。

[0083] DNN模型具有建模较为简单，训练时间较短，以及可以支持流式工作，也就是可以逐帧输入而不需要每次输入一整个序列的优势。

[0084] 需要说明的是，当神经网络映射模型是DNN模型的情况下，本实施例可以通过多种方式选取语音帧集合中的多个语音帧。其中，一种可能的实现方式是将目标语音帧的相邻上下文语音帧作为语音帧集合中的多个语音帧。例如，以目标语音帧为中心，选取相同数量的相邻上文语音帧和相邻下文语音帧，参见图7所示，若窗长为7，目标语音帧为Xt，Xt表示第t个语音帧，则Xt的相邻上下文语音帧包括Xt-3、Xt-2、Xt-1、Xt+1、Xt+2和Xt+3。

[0085] 另一种可能的实现方式是将目标语音帧的间隔上下文语音帧作为语音帧集合中的多个语音帧。本实施例对上下文语音帧的间隔方式不做限定，例如，可以采用倍增选帧法，即按照等比数列的形式倍增地选取上下文语音帧；也可以按照等差数列的形式倍增地选取上下文语音帧等等。参见图8所示，若窗长为7，目标语音帧为Xt，按照等比数列的形式倍增地选取上下文语音帧，则Xt的间隔上下文语音帧包括Xt-4、Xt-2、Xt-1、Xt+1、Xt+2和Xt+4。

[0086] 若神经网络映射模型是LSTM模型或BLSTM模型，二者的输入类似，都可以直接输入

表示一句话的语音帧，而确定表示一句话的语音帧的方式可以是对待处理语音进行语音切分，例如根据待处理语音中的静音片段进行语音切分，得到切分结果。切分结果中切分得到的每个语音片段可以表示一句话，该语音片段中所包括语音帧的PPG可以作为LSTM模型或BLSTM模型的输入。在这种情况下，切分得到的包括目标语音帧的语音片段中的语音帧可以作为语音帧集合，此时，语音帧集合中语音帧的数量是切分得到的包括目标语音帧的语音片段中语音帧的数量，即语音帧集合中语音帧的数量是根据待处理语音的语音切分结果确定的。

[0087] 其中，LSTM模型可以参见图9所示，图9中每个圆圈表示一维参数。当输入语音帧集合中语音帧分别对应的PPG时，便可输出目标语音帧对应的表情参数。LSTM模型的优势是在于能够方便对序列建模，并且能够捕捉到上下文语音帧的信息，但更侧重于上文语音帧的信息。

[0088] BLSTM模型可以参见图10所示，BLSTM模型与LSTM模型类似，区别在于BLSTM中每个隐含层单元能接受序列上下文语音帧两个方向的输入信息，因此，相比LSTM模型的优势是对下文语音帧的信息也能较好地捕捉到，更加适合与上下文语音帧都对表情参数的确定有显著影响的情况。

[0089] 可以理解的是，在根据语音帧集合中语音帧分别对应的语言学信息确定目标语音帧对应的表情参数时，每个目标语音帧对应一个表情参数，多个目标语音帧对应的表情参数之间可能存在突变或衔接不连续的情况。为此，可以对确定出的表情参数进行平滑处理，避免由于表情参数发生突变，进而使得根据表情参数驱动动画形象做出的表情连续性更强，提高动画形象做出表情的真实性。

[0090] 本申请实施例提供两种平滑处理方法，第一种平滑处理方法是均值平滑。

[0091] 由于根据语音帧集合中语音帧分别对应的语言学信息，可以确定出语音帧集合中语音帧分别对应的待定表情参数(即语音帧集合中每个语音帧的表情参数)。由于目标语音帧可以出现在不同的语音帧集合中，从而得到多个目标语音帧的待定表情参数，故，基于目标语音帧在不同语音帧集合中分别确定的待定表情参数可以对目标语音帧的表情参数进行平滑处理，计算目标语音帧对应的表情参数。

[0092] 例如，当目标语音帧为Xt，语音帧集合为{Xt-2、Xt-1、Xt、Xt+1、Xt+2}，确定出该语音帧集合中语音帧分别对应的待定表情参数依次是{Yt-2、Yt-1、Yt、Yt+1、Yt+2}，目标语音帧Xt还可以出现在其它语音帧集合中，例如语音帧集合{Xt-4、Xt-3、Xt-2、Xt-1、Xt}、语音帧集合{Xt-3、Xt-2、Xt-1、Xt、Xt+1}、语音帧集合{Xt-1、Xt、Xt+1、Xt+2、Xt+3}、语音帧集合{Xt、Xt+1、Xt+2、Xt+3、Xt+4}，根据这些语音帧集合，可以确定出目标语音帧在这些集合中分别对应的待定表情参数Yt，即一共得到5个目标语音帧Xt的待定表情参数。将这5个待定表情参数取平均，便可以计算得到目标语音帧Xt对应的表情参数。

[0093] 第二种平滑处理方法是极大似然参数生成算法(Maximum likelihood parameter generation,MLPG)。

[0094] 由于根据语音帧集合中语音帧分别对应的语言学信息可以确定出语音帧集合中目标语音帧对应的待定表情参数(即目标语音帧的表情参数)，同时还可以确定该待定表情参数的一阶差分(或者还有二阶差分)，在给定静态参数(待定表情参数)和一阶差分(或者还有二阶差分)的情况下还原出一个似然最大的序列，通过引入的差分的方式来修正待定

表情参数的变化从而达到平滑的效果。

[0095] 在得到平滑后的表情参数之后，就可以通过自然状态下的动画形象，使得动画形象做出待处理语音对应的表情，通过修改动画形象的参数设置，动画形象做出的表情与待处理语音同步。

[0096] 接下来，将结合具体应用场景，对本申请实施例提供的语音驱动动画方法进行介绍。

[0097] 在该应用场景中，动画形象用于新闻播报，比当待处理语音为“观众朋友们，大家晚上好，欢迎收看今天的新闻联播”，那么，该动画形象需要做出与该待处理语音对应的口型，使得观众感觉该待处理语音就是该动画形象发出的，增强真实感。为此，在获取待处理语音“观众朋友们，大家晚上好，欢迎收看今天的新闻联播”后，可以确定该待处理语音中语音帧对应的语言学信息。针对待处理语音中的每个语音帧例如目标语音帧，根据语言学信息确定语音帧集合中语音帧分别对应的待定表情参数，将目标语音帧在不同语音帧集合中分别确定的待定表情参数取平均，计算目标语音帧对应的表情参数。这样，便可以得到待处理语音中每个语音帧的表情参数，从而驱动动画形象做出“观众朋友们，大家晚上好，欢迎收看今天的新闻联播”对应的口型。

[0098] 基于前述实施例提供的方法，本实施例还提供一种基于人工智能的语音驱动动画装置。参见图11，所述装置包括获取单元1101、第一确定单元1102、第二确定单元1103和驱动单元1104：

[0099] 所述获取单元1101，用于获取待处理语音，所述待处理语音包括多个语音帧；

[0100] 所述第一确定单元1102，用于确定所述待处理语音中语音帧对应的语言学信息，所述语言学信息用于标识所述待处理语音中语音帧所属音素的分布可能性；

[0101] 所述第二确定单元1103，用于根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数；

[0102] 所述驱动单元1104，用于根据所述表情参数驱动动画形象做出对应所述待处理语音的表情。

[0103] 在一种可能的实现方式中，目标语音帧为所述待处理语音中的一个语音帧，针对所述目标语音帧，所述第二确定单元1103，用于：

[0104] 确定所述目标语音帧所处的语音帧集合，所述语音帧集合包括所述目标语音帧和多个语音帧，所述多个语音帧为所述目标语音帧的上下文语音帧；

[0105] 根据所述语音帧集合中语音帧分别对应的语言学信息，确定所述目标语音帧对应的表情参数。

[0106] 在一种可能的实现方式中，所述语音帧集合中语音帧的数量是根据神经网络映射模型确定的，或者，所述语音帧集合中语音帧的数量是根据所述待处理语音的语音切分结果确定。

[0107] 在一种可能的实现方式中，所述多个语音帧为所述目标语音帧的相邻上下文语音帧，或者，所述多个语音帧为所述目标语音帧的间隔上下文语音帧。

[0108] 在一种可能的实现方式中，所述第二确定单元1103，用于：

[0109] 根据所述语音帧集合中语音帧分别对应的语言学信息，确定所述语音帧集合中语音帧分别对应的待定表情参数；

[0110] 根据所述目标语音帧在不同语音帧集合中分别确定的待定表情参数，计算所述目标语音帧对应的表情参数。

[0111] 在一种可能的实现方式中，所述语言学信息包括音素后验概率、瓶颈特征和嵌入特征中任意一种或多种的组合。

[0112] 在一种可能的实现方式中，所述第二确定单元1103，用于：

[0113] 根据所述语言学信息，通过神经网络映射模型确定所述待处理语音中语音帧对应的表情参数；所述神经网络映射模型包括深度神经网络DNN模型、长短期记忆网络LSTM模型或双向长短期记忆网络BLSTM模型。

[0114] 在一种可能的实现方式中，所述第二确定单元1103，用于：

[0115] 根据所述语言学信息和对应的情感向量，确定所述待处理语音中语音帧对应的表情参数。

[0116] 在一种可能的实现方式中，所述第一确定单元1102，用于：

[0117] 确定所述待处理语音中语音帧对应的声学特征；

[0118] 通过自动语音识别模型确定所述声学特征对应的语言学信息。

[0119] 在一种可能的实现方式中，所述自动语音识别模型是根据包括了语音片段和音素对应关系的训练样本训练得到的。

[0120] 本申请实施例还提供了一种设备，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的设备1200，该设备1200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：

[0121] 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。

[0122] 下面结合图12对手机的各个构成部件进行具体的介绍：

[0123] 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。

[0124] 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。

[0125] 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。

[0126] 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。

[0127] 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。

[0128] 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。

[0129] 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。

[0130] 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。

[0131] 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。

[0132] 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。

[0133] 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：

[0134] 获取待处理语音，所述待处理语音包括多个语音帧；

[0135] 确定所述待处理语音中语音帧对应的语言学信息，所述语言学信息用于标识所述待处理语音中语音帧所属音素的分布可能性；

[0136] 根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数；

[0137] 根据所述表情参数驱动动画形象做出对应所述待处理语音的表情。

[0138] 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。

[0139] 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。

[0140] 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。

[0141] 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语音驱动动画方法。

[0142] 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语音驱动动画方法。

[0143] 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。

[0144] 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。

[0145] 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。

[0146] 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。

[0147] 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。

[0148] 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。

[0149] 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。

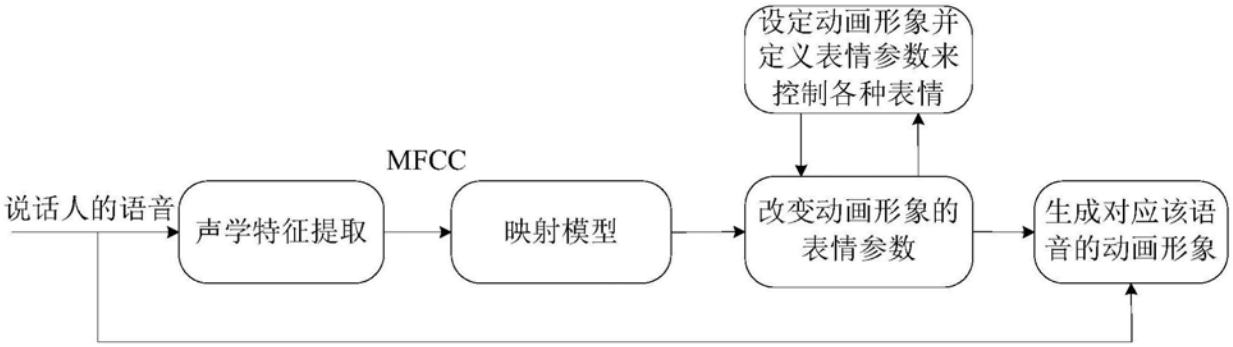

设定动画形象并
定义表情参数来
控制各种表情
MFCC
说话人的语音
声学特征提取
映射模型
改变动画形象的
表情参数
生成对应该语
音的动画形象

图1

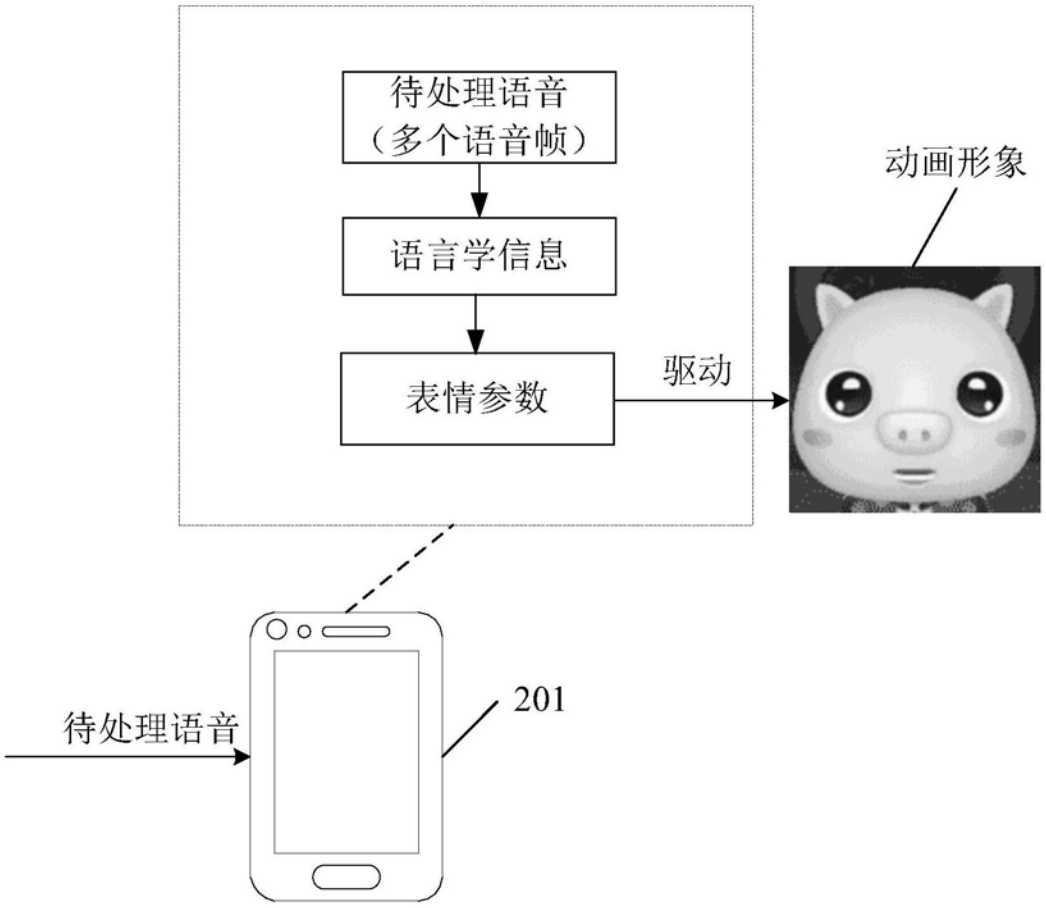

待处理语音
（多个语音帧）
语言学信息
表情参数
驱动
动画形象
待处理语音
201

图2

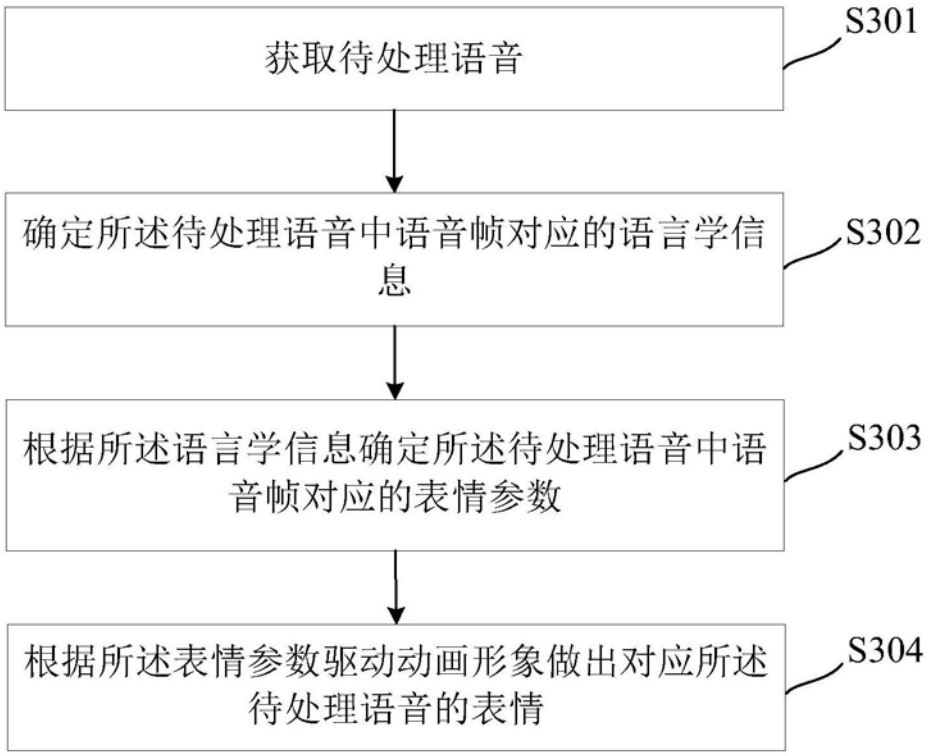

获取待处理语音
S301
确定所述待处理语音中语音帧对应的语言学信息
S302
根据所述语言学信息确定所述待处理语音中语音帧对应的表情参数
S303
根据所述表情参数驱动动画形象做出对应所述待处理语音的表情
S304

图3

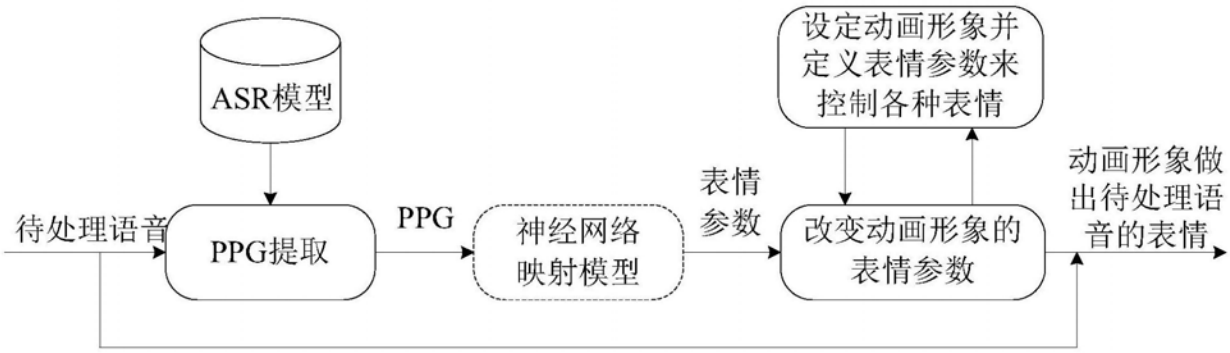

ASR模型
设定动画形象并定义表情参数来控制各种表情
待处理语音
PPG提取
PPG
神经网络映射模型
表情参数
改变动画形象的表情参数
动画形象做出待处理语音的表情

图4

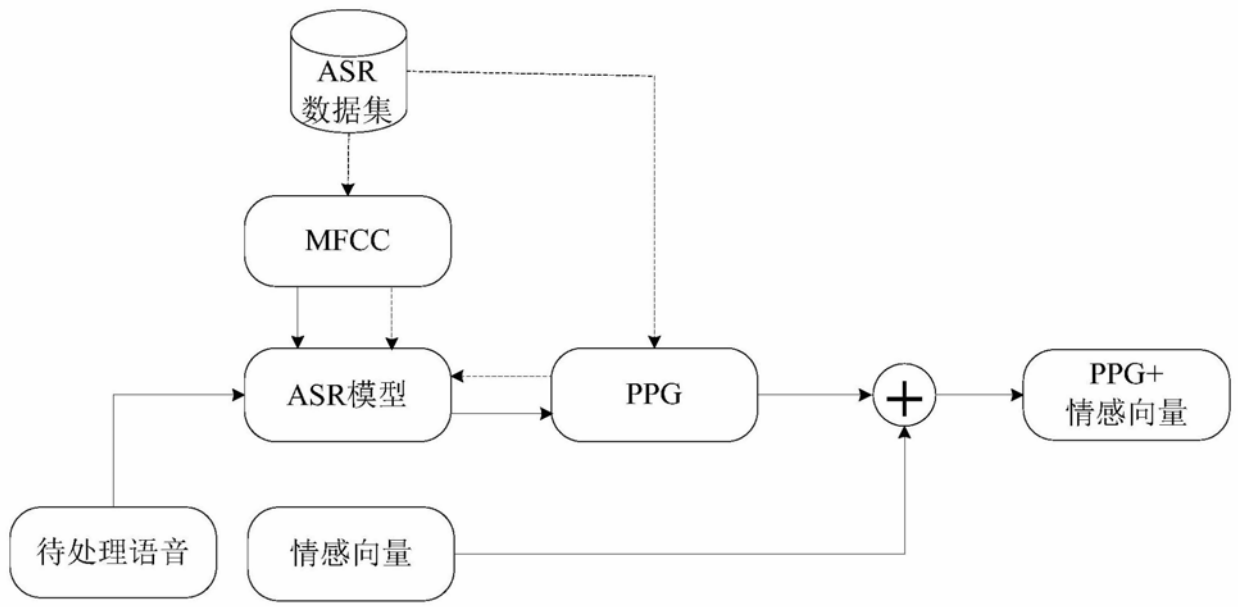

ASR
数据集
MFCC
ASR模型
PPG
PPG+
情感向量
待处理语音
情感向量

图5

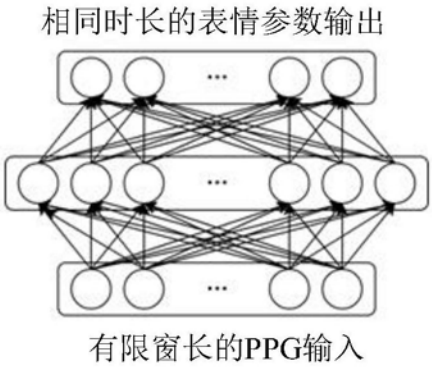

相同时长的表情参数输出
有限窗长的PPG输入

图6

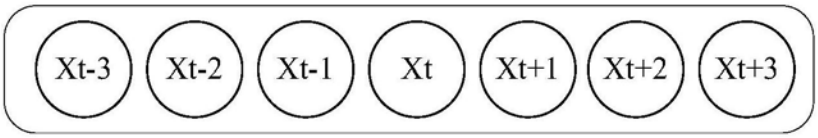

Xt-3
Xt-2
Xt-1
Xt
Xt+1
Xt+2
Xt+3

图7

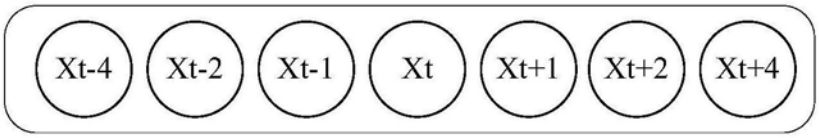

Xt-4
Xt-2
Xt-1
Xt
Xt+1
Xt+2
Xt+4

图8

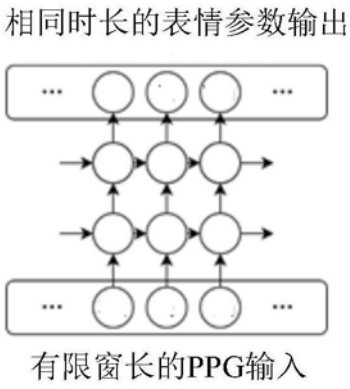

相同时长的表情参数输出
有限窗长的PPG输入

图9

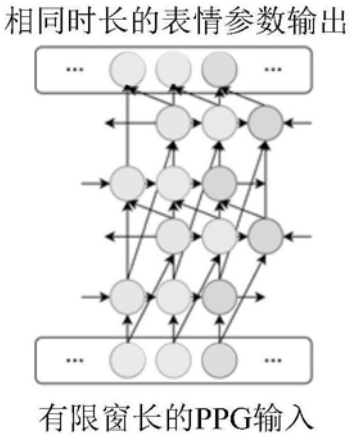

相同时长的表情参数输出
有限窗长的PPG输入

图10

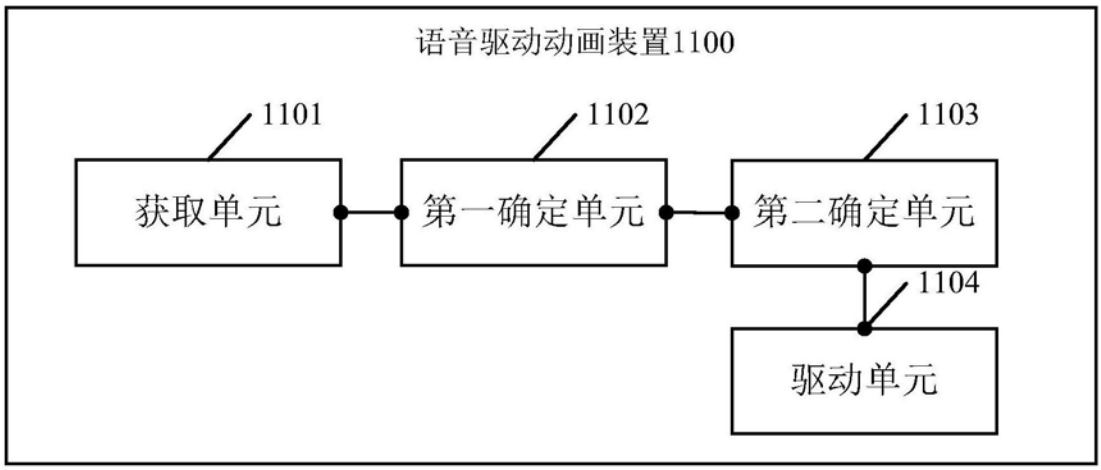

语音驱动动画装置1100
1101
获取单元
1102
第一确定单元
1103
第二确定单元
1104
驱动单元

图11

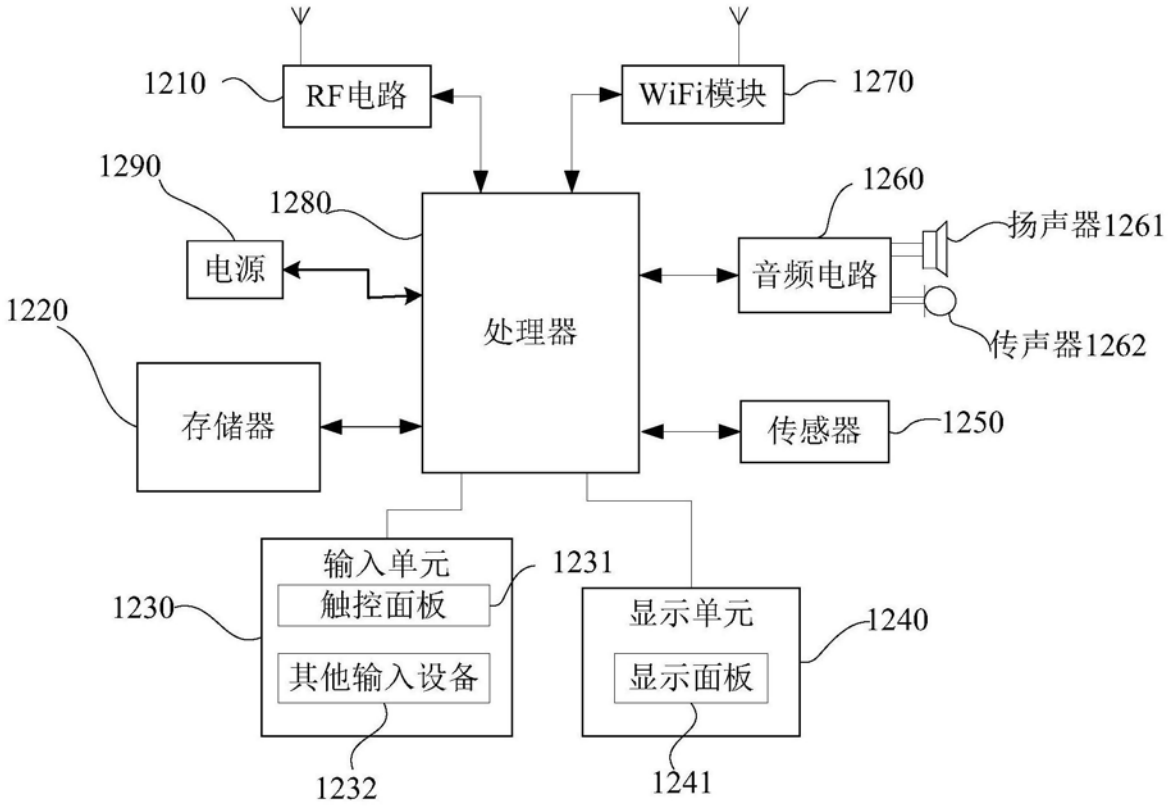

1210
RF电路
WiFi模块
1270
1290
1280
1260
扬声器1261
电源
处理器
音频电路
1220
传声器1262
存储器
传感器
1250
输入单元
1231
触控面板
1230
显示单元
1240
其他输入设备
显示面板
1232
1241

图12

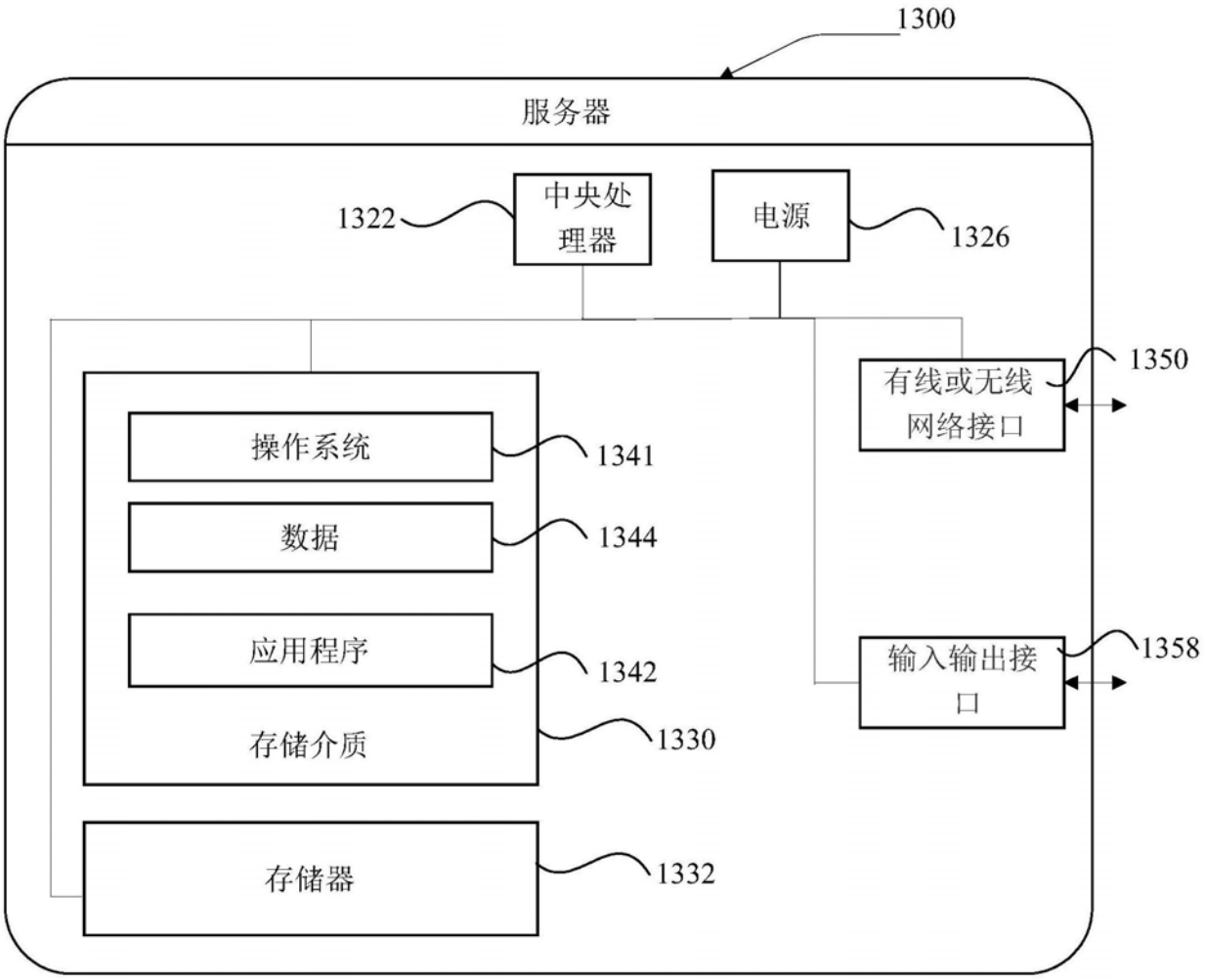

1300
服务器
1322
中央处理器
电源
1326
有线或无线网络接口
1350
操作系统
1341
数据
1344
应用程序
1342
存储介质
1330
输入输出接口
1358
存储器
1332

图13