

视频生成方法、装置和计算机可读存储介质

本申请要求于 2022 年 9 月 16 日提交中国专利局、申请号 202211127727.7、申请名称为“虚拟对象的视频生成方法、装置和计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本申请涉及互联网技术领域，具体涉及视频生成技术。

背景技术

近年来，随着互联网技术的发展，虚拟人等虚拟对象在广告、直播等内容展示场景中的应用越来越广泛。为了提升虚拟对象与真实人类的相似度，往往会控制虚拟对象的口型闭合来模拟真实人类讲话行为。

由于设备、场地以及时间成本等因素的限制，对于需要在短时间内批量生产虚拟对象口型动作的广告场景，一般采用随机控制虚拟对象口型闭合的方法来生成广告场景下虚拟对象模拟讲话的视频。

然而，采用随机控制虚拟对象口型闭合的方法生成的虚拟对象的视频，虚拟对象的口型与音频内容不匹配，虚拟对象的口型生成的准确性较低，使得虚拟对象的视频生成准确性较差，进而导致虚拟对象的视频生成效率较低。

发明内容

本申请实施例提供一种视频生成方法、装置和计算机可读存储介质，可以提高视频数据中虚拟对象的目标部位状态与音频数据相匹配的准确性，进而提高基于虚拟对象的视频生成准确性，进一步提升虚拟对象的视频生成效率。

本申请实施例提供一种视频生成方法，所述方法由计算机设备执行，包括：

获取虚拟对象的初始状态数据以及音频数据；

从所述音频数据提取得到音频特征，并对所述音频数据进行编码，得到编码特征；

将所述音频特征以及所述编码特征进行融合，得到所述虚拟对象的融合音频特征；

25 根据所述融合音频特征和所述初始状态数据得到所述虚拟对象的更新后状态数据；

根据所述更新后状态数据和所述音频数据生成视频数据。

相应的，本申请实施例提供一种视频生成装置，所述装置部署在计算机设备上，包括：

获取单元，用于获取虚拟对象的初始状态数据以及音频数据；

30 提取单元，用于从所述音频数据提取得到音频特征，并对所述音频数据进行编码，得到编码特征；

融合单元，用于将所述音频特征以及所述编码特征进行融合，得到所述虚拟对象的融合音频特征；

预测单元，用于根据所述融合音频特征和所述初始状态数据得到所述虚拟对象的更新后状态数据；

35 生成单元，用于根据所述更新后状态数据和所述音频数据生成视频数据。

此外，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存储有计算机程序，所述计算机程序适于处理器进行加载，以执行本申请实施例所提供的任一种视频生成方法。

此外，本申请实施例还提供一种计算机设备，包括处理器和存储器，所述存储器存储有计算机程序，所述处理器用于运行所述存储器内的计算机程序实现本申请实施例提供的任一种视频生成方法。

本申请实施例还提供一种计算机程序产品，所述计算机程序产品包括计算机程序，所述计算机程序存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取所述计算机程序，处理器执行所述计算机程序，使得所述计算机设备执行本申请实施例提供的任一种视频生成方法。

本申请实施例获取虚拟对象对应的初始状态数据以及音频数据，初始状态数据为虚拟对象的初始外表的数据，为了使得虚拟对象的外表数据与音频数据相匹配，可以在初始状态数据的基础上得到更新后状态数据，从而使得外表数据所表示的虚拟对象的目标部位变化与音频数据相匹配。具体地，可以从音频数据提取得到音频特征，并对音频数据进行编码得到编码特征，从而将音频特征以及编码特征进行融合，得到虚拟对象的融合音频特征。由于该融合音频特征是基于两种不同的特征（即音频特征和编码特征）进行融合得到的，音频特征和编码特征能够从不同的维度体现出音频数据的特征，故融合音频特征可以更加全面准确的表征音频数据的特征。根据这样的融合音频特征和初始状态数据，可以更加准确地得到目标部位的状态基于音频数据发生变化后得到的更新后状态数据，由此根据更新后状态数据和音频数据生成目标部位状态与音频数据相匹配的视频数据。以此，通过在音频数据中提取出音频特征和编码特征，并将音频特征和编码特征进行融合得到融合音频特征，从而根据融合音频特征和初始状态数据来生成表征目标部位的状态变化的更新后状态数据，进而根据更新后状态数据和音频数据生成目标部位状态与音频数据相匹配的视频数据，提高视频数据中目标部位与音频数据匹配的准确性，进而提高虚拟对象的视频生成准确性，进一步提升虚拟对象的视频生成效率。

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，对于本领域技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

- 图 1 是本申请实施例提供了一种视频生成方法实施场景示意图；
- 图 2 是本申请实施例提供了一种视频生成方法的流程示意图；
- 图 3a 是本申请实施例提供的另一种视频生成方法的流程示意图；
- 图 3b 是本申请实施例提供了一种视频生成方法的具体流程示意图；
- 图 3c 是本申请实施例提供了一种从音频数据提取得到音频特征的流程示意图；
- 图 4a 是本申请实施例提供了一种视频生成方法的整体流程示意图；
- 图 4b 是本申请实施例提供了一种确定音频语义特征的整体流程示意图；
- 图 4c 是本申请实施例提供了一种视频生成方法的编码器结构示意图；
- 图 5 是本申请实施例提供了一种视频生成方法的另一具体流程示意图；

图 6 是本申请实施例提供的一种视频生成方法的另一流程示意图；

图 7 是本申请实施例提供的一种视频生成装置的结构示意图；

图 8 是本申请实施例提供的计算机设备的结构示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

本申请实施例提供一种视频生成方法、装置和计算机可读存储介质。其中，该视频生成装置可以集成在计算机设备中，该计算机设备可以是服务器，也可以是终端等设备。

其中，服务器可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、网络加速服务（Content Delivery Network, CDN）、以及大数据和人工智能平台等基础云计算服务的云服务器。终端可以包括但不限于手机、电脑、智能语音交互设备、智能家电、车载终端、飞行器等。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接，本申请在此不做限制。

请参阅图 1，以视频生成装置集成在计算机设备中为例，图 1 为本申请实施例所提供的视频生成方法的实施场景示意图，其中，该计算机设备可以为服务器，也可以为终端，该计算机设备可以获取虚拟对象对应的初始状态数据以及音频数据；从音频数据中提取音频特征，并对音频数据进行编码，得到编码特征；将音频特征以及编码特征进行融合，得到虚拟对象的融合音频特征；根据融合音频特征和初始状态数据得到虚拟对象的更新后状态数据；根据更新后状态数据和音频数据生成视频数据，视频数据中目标部位的状态与音频数据相匹配的。

需要说明的是，本发明实施例可应用于各种场景，包括但不限于云技术、人工智能、智慧交通、辅助驾驶等。图 1 所示的视频生成方法的实施环境场景示意图仅仅是一个示例，本申请实施例描述的视频生成方法的实施环境场景是为了更加清楚的说明本申请实施例的技术方案，并不构成对于本申请实施例提供的技术方案的限定。本领域普通技术人员可知，随着视频生成的演变和新业务场景的出现，本申请提供的技术方案对于类似的技术问题，同样适用。

本申请实施例提供的方案涉及人工智能的机器学习等技术，具体通过如下实施例进行说明。需要说明的是，以下实施例的描述顺序不作为对实施例优选顺序的限定。

本实施例将从视频生成装置的角度进行描述，该视频生成装置具体可以集成在计算机设备中，该计算机设备可以是服务器，本申请在此不作限制。

请参阅图 2，图 2 是本申请实施例提供的视频生成方法的流程示意图。该视频生成方法包括：

在步骤 101 中，获取虚拟对象的初始状态数据以及音频数据。

其中，该虚拟对象可以为采用数字技术在虚拟场景中模拟出来的对象，该虚拟对象可以为采用二维（2D）、三维（3D）模型构建的对象，例如，可以为虚拟人、虚拟动物等各

种虚拟形象，其中，虚拟人可以称为数字人，是指通过计算机模拟出的一个类似真实人类的虚拟形象。该初始状态数据可以为虚拟对象的初始状态的数据，在本申请实施例中，“初始”是相对于后续的更新后状态数据而言，是一个相对而非绝对的概念。

当虚拟对象是 3D 模型构建的对象时，初始状态数据可以是初始状态三维数据。在一种可能的实现方式中，状态可以是指外表，故初始状态数据可以是初始外表数据，相应的，更新后状态数据可以是更新后外表数据。外表数据可以指示虚拟对象的外表，比如可以是虚拟对象的目标区域，如脸部区域，进一步比如可以是该目标区域的目标部位，如口部、眼部等部位的状态，本申请实施例将主要以目标区域是脸部区域为例进行介绍，此时初始状态数据可以是初始脸部数据。这里的状态可以是静态或者动作，如张口、闭口、笑、哭、睁眼、闭眼等。在一种可能的实现方式中，该初始状态数据可以虚拟对象在静态状态下的脸部数据，静态状态可以指虚拟对象未执行动作以及未产生情绪时的状态，例如，可以为虚拟对象在未开口说话、进食、大笑时的脸部数据。该音频数据可以为包括音频信号的数据，能够使得初始脸部的目标部位的状态发生变化，音频数据例如可以为一段广告音频、口播音频等。

参见图 3a 所示，本申请实施例获取初始状态数据和音频数据，对音频数据进行特征提取得到音频特征，以及对音频数据进行编码得到编码特征，将音频特征与编码特征进行特征融合得到融合音频特征，根据融合音频特征和初始状态数据得到更新后状态数据，进而基于更新后状态数据和音频数据渲染成片，得到最终的视频数据。

在本申请实施例中获取音频数据的方式可以包括多种，一种方式是直接获取或采集音频数据，另一种方式，也可以是先获取到文本数据，再根据文本数据来获取到音频数据，其中，该文本数据可以为文本形式的数据，例如，可以为广告文案、新闻文案等文本数据。其中，根据文本数据来获取到音频数据的方式可以有多种，例如，请参考图 3b，图 3b 是本申请实施例提供的一种视频生成方法的具体流程示意图，可以采用语义合成技术(Text To Speech, 简称 TTS) 将文本数据转换为音频数据，比如，可以采用 Tacotron (一个端到端的从文本合成音频的 TTS 模型)，进而对音频数据执行后续步骤的处理，例如对音频数据进行特征提取得到融合音频特征。需要说明的是，基于音频数据得到融合音频特征的过程可以参见图 3a 所示，即对音频数据进行编码得到编码特征，将音频特征与编码特征进行特征融合得到融合音频特征，在图 3b 中对此不再赘述。当初始状态数据是初始脸部数据、且目标部位是口部时，而口部的状态可以由口型参数表征，故为了使得口部的状态与音频数据相匹配，可以根据融合音频特征进行口型参数生成，进而结合口型参数得到更新后脸部数据。基于更新后脸部数据和音频数据渲染成片，得到最终的视频数据。

在步骤 102 中，从音频数据提取得到音频特征，并对音频数据进行编码，得到编码特征。

为了既能从音频数据在人耳听觉角度，又能从虚拟对象的声道特性角度，全面准确的体现音频数据的特征，本申请实施例可以从音频数据提取得到音频特征，并对音频数据进行编码得到编码特征。

音频特征可以是从小人耳听觉角度体现音频数据的特点的特征，人耳对不同频率的声波有不同的听觉敏感度，在本申请实施例中，可以将能够更好地模拟人耳听觉感知的频率称为目标频率。从音频数据提取得到音频特征的方式可以是将音频数据转化到目标频率，然后进行倒谱分析，从而得到音频特征，该音频特征可以称为频率特征。在一种可能的实现方式中，目标频率可以为基于人耳对等距的音高(pitch)变化的感官判断而定的非线性频率，例如，可以为梅尔 (Mel) 频率，相对于赫兹频率，梅尔频率可以代表一般人耳对于频率的感受度。该音频特征可以为表征目标频率下的音频数据的信息，可以为根据人的听觉系统研究成果推导出音频数据在目标频率下的声学特征，例如可以包括频谱特征、振幅、强度、音高 (pitch)、音色 (timbre)、音强 (loudness)、声调 (tone) 等。

其中，音高与声带振动的频率有关，频率越高音高也越高，音高受年龄和性别的影响大。正常人群中，幼儿和女性音高较高，成年男性音高较低。一般用基频来反映音高的高低，基频是声带作周期性振动时一秒钟内声带振动的次数，即声带振动的频率，单位是赫兹(Hz)一般来说，正常儿童的基频在 340H 左右，正常女性的基频在 250Hz 左右，成年男性的基频在 130Hz 左右。另一个反映音高的参数是基频标准差，是基频偏差量的测定值，单位也是 Hz，正常值小于 3Hz。

音色可以是指声音的特色，也可以说是语音的本质，所以又叫做音质。不同的音色是由于声波振动的形式不同形成的，它是一个声音区别于其他声音的基本特征。造成不同音色的条件主要有以下三种：①发音体不同；②发音方法不同；③发音时共鸣器形状不同。

音强就是指声音的强弱，主要取决于发音体的振幅（振幅指发音体振动时最大的位移距离，即发音体振动的幅度）。振幅与音强成正比。振幅大，声音就强；振幅小，声音就弱。音强的单位可以是分贝 (db)，通常情况下，0 分贝：刚能听到的声音；15 分贝：感觉安静；30 分贝：耳语的音量大小；40 分贝：冰箱的嗡嗡声；60 分贝：正常交谈的声音；70 分贝：相当于走在闹市区；85 分贝：汽车穿梭的马路上；95 分贝：摩托车启动声音；100 分贝：装修电钻的声音，等等。

声调可以是指音节在发音过程中的高低抑扬性（音调，或称音高）及顿挫性（韵尾或闭塞音）。声调是依附在音节上的超音段成分，在汉语言中可以有四个声调：平声、上升声、升降声、下降声。

在一种可能的实现方式中，目标频率可以为梅尔频率，则目标频率下的声学特征可以为梅尔倒谱系数 (Mel-frequency cepstral coefficients，简称 MFCC) 特征，此时音频特征是 MFCC 特征。

基于编码的方式得到编码特征可以是模拟人类的发声原理，通过分析声道短管级联的模型得到编码特征的。在一种可能的实现方式中，编码可以是预测编码，预测编码主要是对音频数据进行采样，得到一系列数字，然后利用这些数字之间的组合关系来预测下一个采样值，由此可以通过预测编码提供特征提取的功能，将音频数据转换为计算机设备可以识别的特征向量，即预测编码特征。预测编码特征可以是从小道特性角度表征音频数据的特点的特征，具体可以表征音频数据中音频帧之间的组合关系，即针对音频数据中的音频

帧，可以利用该音频帧的历史音频帧进行组合来拟合该音频帧。该音频帧的历史音频帧为音频数据中位于该音频帧之前的全部或部分音频帧。

在一种可能的实现方式中，预测编码可以是线性预测编码，此时的预测编码特征可以为线性预测编码特征（是一种线性特征），从而在前述音频特征为 MFCC 特征时，可以结合 MFCC 特征，从线性特征和非线性特征（即 MFCC 特征）两方面考虑音频数据的特点。

线性预测编码（Linear Predictive Coding，简称 LPC）的主要思路是通过对音频数据进行采样，得到一系列数字，然后利用这些数字之间的线性关系来预测下一个采样值，由此可以通过线性预测编码提供特征提取的功能，将音频数据转换为计算机设备可以识别的特征向量，即线性预测编码特征。该线性预测编码特征可以是对音频数据进行线性预测编码得到的特征，该线性预测编码特征可以表征音频数据中音频帧之间的线性组合关系，是一种线性特征。通常情况下，音频数据中可以包括多个音频帧，对于除了第一个音频帧之外的其他音频帧，其他音频帧中每个音频帧都可以存在对应的历史音频帧，位于该音频帧之前的全部或部分音频帧都可以作为该音频帧的历史音频帧。在本申请实施例中，针对一个音频帧，可以获取该音频帧的历史音频帧，将该音频帧的历史音频帧进行线性组合，从而利用历史音频帧之间的线性组合关系拟合该音频帧，从而获得该音频帧与历史音频帧之间的线性组合关系。

在一种可能的实现方式中，从音频数据提取得到音频特征的方式可以有多种，例如，可以将该音频数据由当前频率转换为目标频率，并基于目标频率的音频数据，确定该音频数据对应的目标频谱，根据该目标频谱对该目标频率的音频数据进行倒谱分析，得到该音频数据对应的音频特征。

其中，该当前频率可以为当前音频数据中音频信号的频率，例如，可以为赫兹（Hz）频率，该目标频率可以为梅尔频率，该目标频谱可以为根据目标频率在音频数据中提取出的表征音频数据的音频信号变化的参数，例如，当目标频率可以为梅尔频率时，目标频谱可以为音频数据的梅尔频谱。

倒谱（cepstrum）可以是指音频数据的傅里叶变换谱经对数运算后再进行的傅里叶反变换，倒谱分析即先对音频数据的傅里叶变换谱进行对数运算，再对对数运算后的结果进行傅里叶反变换。

其中，将该音频数据由当前频率转换为目标频率，并基于目标频率的音频数据，确定该音频数据对应的目标频谱的方式可以有多种，例如，可以将该音频数据进行分帧处理，得到多个音频帧，针对多个音频帧中的每个音频帧，对该音频帧进行加窗处理，并对加窗处理后的音频帧进行频域转换处理，得到该音频帧的频谱信息，对该频谱信息进行目标频率的滤波处理，得到该音频数据对应的目标频谱。

其中，该音频帧可以为对音频数据进行划分后得到的音频数据帧，该频谱信息可以为表征音频数据在频域上的信号分布的信息。

其中，将该音频数据进行分帧处理的方式可以有多种，例如，可以先对音频数据进行预加重处理，以提升高频部分，使音频信号的频谱变得平坦，同时，可以消除音频数据中声带和嘴唇造成的效应，补偿语音信号受到发音系统所压抑的高频部分等，从而可以对预

加重处理后的音频数据进行分帧处理，将音频数据划分为多个音频帧，以便于对音频数据进行分析。

在将该音频数据进行分帧处理之后，由于计算机设备处理的信号长度是有限的，因此，为了更好的对音频帧执行后续步骤，便可以对该音频帧进行加窗处理，从而将音频帧截断成多个小块（即窗口）进行处理。在信号处理技术中，加窗处理是一种基本技术，它可以用于改善信号的质量和性能。加窗处理也可以称为窗口函数处理，窗口函数用于将信号分割成等宽的小块，这些小块被称为窗口，每个窗口有一个窗口函数。加窗处理的目的是确定一种合适的窗口函数，以便窗口的大小可以合理地改变，以满足信号处理的要求。这样，可以有效地改善信号的传输效率，减少噪声，从而提高信号的质量。

对音频帧进行加窗处理的过程主要分为几个步骤：第一步是将音频帧分割成若干小块，即窗口；第二步是为每个窗口设置一个适当的窗口函数；第三步是按照指定的窗口函数对音频帧进行加窗处理；最后一步是将处理后的音频帧重新组合成初始的音频帧。其中，对该音频帧进行加窗处理的方式可以有多种，例如，可以采用窗函数对音频帧进行加窗处理，例如，可以采用汉明窗来对音频帧进行加窗处理。

在对该音频帧进行加窗处理之后，便可以对加窗处理后的音频帧进行频域转换处理。频域转换处理用于将时域信号转换为频域信号，在本申请实施例中，主要是指将加窗处理后的音频帧在时域上的信号转换为频域上的能量分布。其中，对加窗处理后的音频帧进行频域转换处理的方式可以有多种，例如，可以采用快速傅里叶变换（Fast Fourier Transform，简称 FFT）等傅里叶变换方法来将加窗处理后的音频帧在时域上的信号转换为频域上的能量分布，以此得到每一音频帧对应的频谱信息。

在对加窗处理后的音频帧进行频域转换处理之后，便可以对该频谱信息进行目标频率的滤波处理。其中，滤波处理可以是指滤除频谱信息中的冗余信息，对该频谱信息进行目标频率的滤波处理的方式可以有多种，比如，可以采用目标频率对应的滤波器组对音频帧的频谱信息进行目标频率的滤波处理，从而可以去除频谱信息中的冗余信息，得到该音频数据对应的目标频谱。该滤波器组可以为基于将目标频率的频谱按人耳敏感程度进行划分得到的多个滤波器组，譬如，可以为梅尔滤波器组。

基于上述介绍，从音频数据提取得到音频特征的流程示意图可以参见图 3c 所示，将音频数据进行分帧处理，得到多个音频帧；针对多个音频帧中的每个音频帧，对音频帧进行加窗处理，并对加窗处理后的音频帧进行频域转换处理，得到音频帧的频谱信息；对频谱信息进行目标频率的滤波处理，得到音频数据对应的目标频谱，根据目标频谱对目标频率的音频数据进行倒谱分析，得到音频数据对应的音频特征。

在将该音频数据由当前频率转换为目标频率，并基于目标频率的音频数据，确定该音频数据对应的目标频谱之后，可以根据该目标频谱对该目标频率的音频数据进行倒谱分析。其中，根据该目标频谱对该目标频率的音频数据进行倒谱分析的方式可以有多种，例如，可以将该目标频谱进行非线性转换，得到非线性转换结果，对该非线性转换结果进行低频信息提取，得到该音频数据的低频信息，在该低频信息中提取出音频数据对应的音频特征。

其中，该非线性转换结果可以为对目标频谱进行非线性转换后的结果，该低频信息可以以为将音频数据的能量集中在低频率中而得到的信息，该低频信息可以是目标频率倒谱系数，目标频率倒谱系数可以为组成目标频率倒谱的系数，例如，可以为梅尔频率倒谱系数。在一种可能的实现方式中，频率的高低可以通过频率的范围进行界定，通常情况下，频率在 50 赫兹以下可以称为极低频的频段，频率在 50 赫兹-200 赫兹之间可以称为低频的频段，频率在 200 赫兹到 6000 赫兹可以称为中间频段，频率在 6000 赫兹-20000 赫兹可以称为高频的频段。将音频数据的能量集中在低频的频段中得到的信息可以称为低频信息。

其中，将该目标频谱进行非线性转换的方式可以有多种，例如，可以将目标频谱进行对数运算，得到非线性转换结果，由于人耳对声音的感知并不是线性的，因此，将目标频谱进行类型对数运算这种非线性关系的转换可以更好的实现对音频数据的音频信号的描述。

目标频谱可以表示为：

$$X[k]=H[k]E[k]$$

其中， $X[k]$ 可以表示目标频谱， $H[k]$ 可以表示包络， $E[k]$ 可以表示目标频谱的细节。只考虑幅度时，目标频谱可以表示为：

$$|X[k]|=|H[k]||E[k]|$$

两边取对数，即对目标频谱进行对数运算：

$$\log|X[k]|=\log|H[k]|+\log|E[k]|$$

在将该目标频谱进行非线性转换之后，便可以对非线性转换结果进行低频信息提取。其中，对该非线性转换结果进行低频信息提取的方式可以有多种，例如，可以采用离散余弦变换 (DCT for Discrete Cosine Transform)、离散傅里叶变换(DFT for Discrete Fourier Transform)等变换方法对该非线性转换结果进行低频信息提取，以将音频数据对应的能量集中到低频部分。

在该非线性转换结果进行低频信息提取之后，便可以在该低频信息中提取出音频数据对应的音频特征。其中，在该低频信息中提取出音频数据对应的音频特征的方式可以有多种，例如，可以在该低频信息中选取部分系数作为音频特征，比如，假设该目标频率为梅尔频率，则可以在离散余弦变换后得到的低频信息中将第 2 个到第 13 个系数作为梅尔频率倒谱系数，从而可以将每一音频帧对应的梅尔频率倒谱系数作为该音频帧对应的特征信息，根据每一音频帧的特征信息，则可以得到音频数据对应的音频特征。在一种可能的实现方式中，由于音频信号是时域连续的，分帧提取的音频帧的特征信息只反应了本帧音频的特性，为了使特征更能体现出时域连续性，可以将音频数据对应的音频帧对应的特征信息在特征维度增加前后帧信息的维度，从而得到音频数据对应的音频特征。

其中，对该音频数据进行编码的方式可以有多种，在一种可能的实现方式中，该编码可以是线性编码。此时对音频数据进行编码，得到编码特征的方式可以是：针对音频数据中包括的每个音频帧，可以根据该音频帧以及该音频帧在该音频数据中对应的历史音频帧，构建该音频帧和该历史音频帧之间的线性组合关系，基于该线性组合关系中，确定该音频数据对应的编码特征。此时，该编码特征可以是线性预测编码特征。

其中，该音频帧可以为对音频数据进行分帧处理得到的音频数据帧。该历史音频帧可以为在音频数据对应的时间序列中处于当前音频帧之前的音频帧，例如，假设音频数据为(Y1、Y2、Y3、Y4)，其中，Y1、Y2、Y3、Y4为音频数据中以时间先后顺序排列的音频帧，则对于音频帧Y2，Y1可以为Y2的历史音频帧，对于音频帧Y3，Y1以及Y2可以为Y3的历史音频帧。该线性组合关系可以为音频帧与其对应的历史音频帧之间构成的线性组合形式的关系。

其中，该线性预测编码特征可以为LPC特征，由于一个语音的采样可以用过去若干个语音采样的线性组合来逼近，通过使线性预测到的采样逼近实际语音采样，从而可以求取一组唯一的预测系数。这里的预测系数就是线性组合中所用的加权系数，也就是线性预测编码特征。

其中，根据该音频帧以及该音频帧在该音频数据中对应的历史音频帧，构建该音频帧和该历史音频帧之间的线性组合关系的方式可以有多种，例如，假设音频数据为(Y1、Y2、Y3、Y4)，对于音频帧Y3，可以构建该音频帧和该历史音频帧之间的线性组合关系为 $Y3=a1*Y1+a2*Y2$ ，其中，a1、a2可以为该线性组合中的加权系数。

在根据该音频数据的音频帧以及该音频帧在该音频数据中对应的历史音频帧，构建该音频帧和该历史音频帧之间的线性组合关系之后，便可以基于该线性组合关系中，确定该音频数据对应的编码特征。其中，基于该线性组合关系中，确定该音频数据对应的编码特征的方式可以有多种，比如，可以对该线性组合关系进行收敛，使得预测的音频信号与初始的音频信号误差最小，从而可以求解出加权系数，从而可以根据音频数据中每一音频帧对应的加权系数得到音频数据对应的编码特征。例如，可以采用最小均方误差等均方误差准则来对该线性组合关系进行收敛，从而可以求解出加权系数。

在步骤103中，将音频特征以及编码特征进行融合，得到虚拟对象的融合音频特征。

其中，特征融合可以是指在给定不同属性的特征时，将不同属性的特征组合到一起从而得到新的特征。特征融合可以帮助捕捉到不同属性的特征之间的关系，从多个角度去学习和利用数据中的信息，从而增强特征的表达能力。在本申请实施例中，不同属性的特征可以是指音频特征以及编码特征，将音频特征以及编码特征组合到一起得到融合特征的过程可以称为特征融合。该融合音频特征可以为融合音频特征以及编码特征后得到的音频特征。

其中，将音频特征以及编码特征进行融合的方式可以有多种，例如，可以将音频特征以及编码特征进行拼接，从而可以得到虚拟对象的融合音频特征。

在一实施例中，可以采用预设特征权重对该音频特征以及该编码特征进行加权，将加权后的音频特征以及加权后的编码特征进行拼接处理，得到该虚拟对象的融合音频特征。

其中，该预设特征权重可以为预先为音频特征（比如频率特征，具体可以是MFCC特征）以及该编码特征（比如线性编码特征）设定的一个权重值，用于实现音频数据对应的音频特征和编码特征的融合。例如，假设音频特征M对应的预设特征权重为m，编码特征N对应的预设特征权重为n，则可以采用预设特征权重对该音频特征以及该编码特征进行

加权, 得到加权后的音频特征 $m*M$ 以及该编码特征 $n*N$, 从而可以将加权后的音频特征以及加权后的编码特征进行拼接处理, 得到该虚拟对象的融合音频特征 $m*M+n*N$ 。

当音频特征是频率特征, 编码特征是线性预测编码特征时, 特征融合方式与上述类似, 此处不再赘述。

- 5 在本申请实施例中, 预设特征权重可以是基于特征的重要程度进行设置的, 重要程度越高的特征, 预设特征权重也越高, 重要程度越低的特征, 预设特征权重也越低。通常情况下, 每个特征的预设特征权重是大于 0 且小于 1 的数值。在一种可能的实现方式中, 音频特征的预设特征权重与编码特征的预设特征权重之和可以为 1。

在步骤 104 中, 根据融合音频特征和初始状态数据得到虚拟对象的更新后状态数据。

- 10 为了使得虚拟对象的状态数据与音频数据相匹配, 在本申请实施例中, 可以基于音频数据对初始状态数据进行更新, 从而使得更新后状态数据中目标部位的状态与音频数据同步。故在本申请实施例中, 可以基于融合音频特征和初始状态数据得到虚拟对象的更新后状态数据。其中, 外表数据可以为虚拟对象的外表的数据, 该目标部位可以为虚拟对象的外表
15 中需要根据音频数据进行更新的部位, 例如, 可以为嘴巴, 以此可以根据音频数据的内容将虚拟对象的口型进行适配变化, 实现自然、真实的展示虚拟对象的讲话过程。在一些可能的实现方式中, 目标部位还可以包括眼睛、脸颊、额头等部位, 可以在虚拟对象基
20 于音频数据进行讲话时, 实现嘴巴(口型)、眼睛、脸颊、额头等部位的变化与音频数据中的内容相匹配, 实现真实的虚拟对象讲话场景。该更新后状态数据可以为基于融合音频特征得到的虚拟对象的初始状态数据随着音频数据中内容发生变化得到的状态数据, 可以
25 为表征虚拟对象的外表目标部位的状态基于音频数据发生变化后的状态数据。

- 其中, 根据融合音频特征和初始状态数据得到更新后状态数据的方式可以有多种, 在一种可能的实现方式中, 可以采用训练好的神经网络模型, 根据融合音频特征和初始状态数据得到更新后状态数据。其中, 该神经网络模型可以为卷积神经网络(Convolutional Neural Networks, 简称 CNN)、循环神经网络(Recurrent Neural Network, 简称 RNN)、长短期
30 记忆网络(LSTM, Long Short-Term Memory)、通用不确定性推理模型(GUR)及基于双向长短时记忆的神经网络模型(Bi-directional Long Short-Term Memory, 简称 BiLSTM)等模型。例如, 请参考图 4a, 图 4a 是本申请实施例提供的一种视频生成方法的整体流程示意图, 以 RNN 模型为例, 可以采用 RNN 模型根据融合音频特征和初始状态数据对虚拟对象
35 的外表数据进行更新。其中, RNN 模型中可以包括编码器结构(Encoder)以及解码器结构(Decoder)。在通过 RNN 模型根据融合音频特征和初始状态数据得到更新后状态数据时, 视频生成的具体过程可以是基于音频数据得到的融合音频特征输入至编码器结构, 通过编码器结构从融合音频特征中提取音频语义特征。将音频语义特征和初始外表特征输入至解码器结构, 由解码器结构输出更新后状态数据, 进而基于更新后状态数据和音频数据生成视频数据。

- 在另一种可能的实现方式中, 根据融合音频特征和初始状态数据得到虚拟对象更新后状态数据的方式可以是对该融合音频特征进行语义特征提取, 得到该音频数据对应的音频语义特征, 根据该初始状态数据以及该音频语义特征, 生成虚拟对象的更新后状态数据。

其中，该音频语义特征可以为表征音频数据的语义信息的特征。例如，请继续参考图 4a，可以采用编码器结构对该融合音频特征进行语义特征提取，得到该音频数据对应的音频语义特征，采用解码器结构根据该初始状态数据以及该音频语义特征，生成表征该目标部位的状态变化的更新后状态数据。

5 通过对音频语义特征进行提取，可以更好的理解音频数据的语义，不同语义可能导致与音频数据相匹配的目标部位的状态不同，进而提高更新后状态数据的准确性。

其中，对该融合音频特征进行语义特征提取的方式可以有多种，例如，可以将该融合音频特征进行时间间隔划分，得到每一时间间隔的音频子特征，按照时间间隔的顺序，分别获取时间间隔的音频子特征对应的历史隐藏特征，根据该历史隐藏特征，对该音频子特征进行特征编码，得到该音频子特征对应的语义特征，并基于该语义特征，确定该融合音频特征对应的音频语义特征，参见图 4b 所示。

10 其中，该时间间隔 (timestep_size) 可以为表征处理融合音频特征的单位，该音频子特征可以为以时间间隔为划分标准对融合音频特征进行划分后得到的子特征。该历史隐藏特征可以为基于历史音频子特征生成的隐藏特征，该历史音频子特征可以为时间间隔的上一时间间隔的音频子特征。该隐藏特征可以表征当前输入的音频子特征与上一时间间隔输出的状态数据之间的关联关系，例如，可以为神经网络模型中的隐藏层，该隐藏特征可以每一时间间隔对应的隐藏层中的信息。该语义特征可以为表征每一音频子特征的语义信息的特征。在一种可能的实现方式中，将融合音频特征对应的每一音频子特征对应的语义特征进行拼接，就可以得到音频数据对应的音频语义特征。

20 其中，根据该历史隐藏特征，对该音频子特征进行特征编码，得到该音频子特征对应的语义特征的方式可以有多种，例如，请继续参考图 4a，以 RNN 模型为例，可以采用 RNN 模型中的编码器结构对融合音频特征进行语义特征提取。具体的，可以参考图 4c，图 4c 是本申请实施例提供的一种视频生成方法的编码器结构示意图，其中，h0、h1、h2、h3 可以表示编码器结构中隐藏层中的隐藏特征，x1、x2、x3 表示每一时间间隔输入的音频子特征，C 表示特征编码得到的音频语义特征。每一时间间隔对应的隐藏特征可以根据上一时间间隔的隐藏特征以及该时间间隔输入的音频子特征来确定。假设该时间间隔的输入为音频子特征 x2，则音频子特征 x2 对应的历史隐藏特征为 h1，从而可以采用编码器根据该历史隐藏特征 h1，对该音频子特征 x2 进行特征编码，得到该音频子特征对应的语义特征 h2，也即该时间间隔对应的隐藏特征。

30 在该音频子特征进行特征编码之后，便可以基于该语义特征，确定该融合音频特征对应的音频语义特征。其中，基于该语义特征，确定该融合音频特征对应的音频语义特征的方式可以有多种，例如，可以将融合音频特征的最后一个时间间隔的音频子特征对应的隐藏特征（也是语义特征）作为该融合音频特征对应的音频语义特征。例如，请继续参考图 4c，可以将 x3 对应的语义特征 h3 作为音频语义特征 C。

35 基于隐藏特征确定音频语义特征，从而可以考率到当前输入的音频子特征与上一时间间隔输出的状态数据之间的关联关系，进而提高音频语义特征的准确性，以便后续得到更加准确的更新后状态数据。

在对该融合音频特征进行语义特征提取之后，便可以根据该初始状态数据以及该音频语义特征，生成更新后状态数据。其中，根据该初始状态数据以及该音频语义特征，生成更新后状态数据的方式可以有多种，例如，可以获取上一时间间隔生成的历史状态数据，根据该历史状态数据以及该音频语义特征，生成当前时间间隔对应的当前状态数据，基于该当前状态数据，确定更新后状态数据。

其中，该历史状态数据可以为在该上一时间间隔，基于该初始状态数据生成的状态数据，即可以为当前时间间隔对应的上一时间间隔的状态数据。该当前状态数据可以为根据历史状态数据以及音频语义特征生成的状态数据，即当前时间间隔的状态数据。

其中，根据该历史状态数据以及该音频语义特征，生成当前时间间隔对应的当前状态数据的方式可以有多种，例如，可以在该历史状态数据中识别出目标部位参数，基于该音频语义特征，对该目标部位参数进行调整，根据调整后的目标部位参数，生成当前时间间隔对应的当前状态数据。

其中，该目标部位参数可以为该历史状态数据中表征该虚拟对象的目标部位的参数，例如，可以为历史状态数据中表征虚拟对象的嘴巴的参数。

本申请的主要目的是使得目标部位与音频数据相匹配，而目标部位的状态可以由目标部位参数来表征，故上述方式可以直接基于音频语义特征调整目标部位参数，进而得到当前状态数据，从而仅通过调整目标部位参数便可以简单、直接、准确地得到当前状态数据。

在步骤 105 中，根据更新后状态数据和音频数据生成视频数据。

更新后状态数据中目标部位的状态是按照音频数据进行变化的，不过更新后状态数据是虚拟对象在画面上的体现，为了得到虚拟对象的视频数据，还需要将更新后状态数据与音频数据进行叠加。在叠加的过程中，需要保证目标部位的状态与音频数据相匹配，从而实现目标部位的状态与音频数据同步。其中，该视频数据可以为基于更新后状态数据和音频数据生成的视频，可以为包括虚拟对象的目标部位状态与音频数据相匹配的视频，例如，可以为包括虚拟对象的口型的变化与音频数据相匹配的视频。

其中，根据该更新后状态数据和音频数据生成视频数据的方式可以有多种。在一种可能的实现方式中，为了避免音频数据的开始时间和更新后状态数据的开始时间不一致，或者音频数据的结束时间和更新后状态数据的结束时间不一致等问题，在生成视频数据时，可以基于该更新后状态数据对该虚拟对象进行渲染，得到该虚拟对象的图像数据，对该更新后状态数据与该音频数据进行时间对齐，从而使得该图像数据与该音频数据同步，基于时间对齐后的图像数据和音频数据，生成视频数据。

这里的时间对齐可以是指使得音频数据的开始时间和更新后状态数据的开始时间一致，并且音频数据的结束时间和更新后状态数据的结束时间一致，从而使得视频数据中目标部位的状态与音频数据能够有更好的匹配效果。

当初始状态数据是初始脸部三维数据时，该图像数据可以为根据更新后状态数据对虚拟对象的三维模型进行渲染后得到数据，例如，可以只包括虚拟对象的脸部，也可以包括虚拟对象的整个身体结构的图像等，可以根据具体情况进行设计。该音频数据可以为时间

序列，基于该时间序列的音频数据预测得到的更新后状态数据也是时间序列，因此，音频数据与更新后状态数据之间可以进行时间对齐。

其中，基于该更新后状态数据对该虚拟对象进行渲染的方式可以有多种，例如，可以采用渲染引擎（UE）以及三维图形图像软件（Blender）等渲染软件来对虚拟对象的三维模型进行渲染。

在一实施例中，可以将本申请实施例提供的视频生成方法应用在广告场景中，例如，以该虚拟对象为虚拟人为例，请参考图 5，图 5 是本申请实施例提供的一种视频生成方法的另一具体流程示意图，其中，本申请实施例提供的视频生成方法可以集成在音频驱动模块中，以此，可以上传广告的图片，并选择虚拟人形象，进而可以上传音频数据或者文本数据，在上传文本数据时，可以通过语音合成技术将文本数据转换为音频数据，在上传音频数据时，可以直接通过音频驱动模块获取更新后状态数据，并根据更新后状态数据渲染为视频数据。以此，实现了在低成本下，实现批量音频驱动 3D 虚拟人的口型参数生成，并且保证了生成的虚拟人的口型与音频数据的匹配性。

由以上可知，本申请实施例获取虚拟对象对应的初始状态数据以及音频数据，初始状态数据为虚拟对象的初始外表的数据，为了使得虚拟对象的外表数据与音频数据相匹配，可以在初始状态数据的基础上得到更新后状态数据，从而使得外表数据所表示的虚拟对象的目标部位变化与音频数据相匹配。具体地，可以从音频数据提取得到音频特征，并对音频数据进行编码得到编码特征，从而将音频特征以及编码特征进行融合，得到虚拟对象的融合音频特征。由于该融合音频特征是基于两种不同的特征（即音频特征和编码特征）进行融合得到的，音频特征和编码特征能够从不同的维度体现出音频数据的特征，故融合音频特征可以更加全面准确的表征音频数据的特征。根据这样的融合音频特征和初始状态数据，可以更加准确地得到目标部位的状态基于音频数据发生变化后得到的更新后状态数据，由此根据更新后状态数据和音频数据生成目标部位状态与音频数据相匹配的视频数据。以此，通过在音频数据中提取出音频特征和编码特征，并将音频特征和编码特征进行融合得到融合音频特征，从而根据融合音频特征和初始状态数据来得到更新后状态数据，进而根据更新后状态数据和音频数据生成目标部位状态与音频数据相匹配的视频数据，提高视频数据中目标部位与音频数据匹配的准确性，进而提高虚拟对象的视频生成准确性，进一步提升虚拟对象的视频生成效率。

根据上面实施例所描述的方法，以下将举例作进一步详细说明。

在本实施例中，将以该视频生成装置具体集成在计算机设备为例进行说明。其中，以服务器执行该视频生成方法、并以该虚拟对象为虚拟人、该目标部位为嘴巴为例进行具体的描述。

为了更好的描述本申请实施例，请参阅图 6，图 6 为本申请实施例提供的视频生成方法的另一流程示意图。具体流程如下：

在步骤 201 中，服务器获取虚拟对象的初始状态数据以及音频数据，将音频数据由当前频率转换为目标频率，基于目标频率的音频数据，确定音频数据对应的目标频谱，根据目标频谱对目标频率的音频数据进行倒谱分析，得到音频数据对应的音频特征。

其中，也可以获取到文本数据，根据文本数据来获取到音频数据，其中，该文本数据可以为文本形式的数据，例如，可以为广告文案、新闻文案等文本数据。其中，根据文本数据来获取到音频数据的方式可以有多种，例如，请参考图 3，可以采用语义合成技术（将文本数据转换为音频数据，比如，可以采用 Tacotron 模型。

5 服务器在获取虚拟对象对应的初始状态数据以及音频数据之后，便可以将音频数据由当前频率转换为目标频率，基于目标频率的音频数据，确定音频数据对应的目标频谱。其中，服务器将该音频数据由当前频率转换为目标频率，并基于目标频率的音频数据，确定该音频数据对应的目标频谱的方式可以有多种，例如，服务器可以将该音频数据进行分帧处理，得到多个音频帧，针对多个音频帧中的每个音频帧，对该音频帧进行加窗处理，并
10 对加窗处理后的音频帧进行频域转换处理，得到该音频帧的频谱信息，对该频谱信息进行目标频率的滤波处理，得到该音频数据对应的目标频谱。

其中，服务器将该音频数据进行分帧处理的方式可以有多种，例如，服务器可以先对音频数据进行预加重处理，以提升高频部分，使音频信号的频谱变得平坦，同时，可以消除音频数据中声带和嘴唇造成的效应，补偿语音信号受到发音系统所压抑的高频部分等，
15 从而可以对预加重处理后的音频数据进行分帧处理，将音频数据划分为多个音频帧，以便于对音频数据进行分析。

服务器在将该音频数据进行分帧处理之后，便可以对该音频帧进行加窗处理。其中，服务器对该音频帧进行加窗处理的方式可以有多种，例如，服务器可以采用窗函数对音频帧进行加窗处理，例如，可以采用汉明窗来对音频帧进行加窗处理。

20 服务器在对该音频帧进行加窗处理之后，便可以对加窗处理后的音频帧进行频域转换处理。其中，服务器对加窗处理后的音频帧进行频域转换处理的方式可以有多种，例如，服务器可以采用快速傅里叶变换（FFT）等傅里叶变换方法来将加窗处理后的音频帧在时域上的信号转换为频域上的能量分布，以此得到每一音频帧对应的频谱信息。

服务器在对加窗处理后的音频帧进行频域转换处理之后，便可以对该频谱信息进行目标频率的滤波处理。其中，服务器对该频谱信息进行目标频率的滤波处理的方式可以有多种，比如，服务器可以采用目标频率对应的滤波器组对音频帧的频谱信息进行目标频率的滤波处理，从而可以去除频谱信息中的冗余信息，得到该音频数据对应的目标频谱。该滤波器组可以为基于将目标频率的频谱按人耳敏感程度进行划分得到的多个滤波器组，譬如，可以为梅尔滤波器组。

30 其中，服务器根据该目标频谱对该目标频率的音频数据进行倒谱分析的方式可以有多种，例如，服务器可以将该目标频谱进行非线性转换，得到非线性转换结果，对该非线性转换结果进行低频信息提取，得到该音频数据的低频信息，在该低频信息中提取出音频数据对应的音频特征。

其中，服务器将该目标频谱进行非线性转换的方式可以有多种，例如，服务器可以将目标频谱进行对数运算，得到非线性转换结果，由于人耳对声音的感知并不是线性的，因此，将目标频谱进行类型对数运算这种非线性关系的转换可以更好的实现对音频数据的音频信号的描述。

服务器在将该目标频谱进行非线性转换之后，便可以对该非线性转换结果进行低频信息提取。其中，服务器对该非线性转换结果进行低频信息提取的方式可以有多种，例如，可以采用离散余弦变换、离散傅里叶变换等变换方法对该非线性转换结果进行低频信息提取，以将音频数据对应的能量会集中到低频部分。

5 服务器在对该非线性转换结果进行低频信息提取之后，便可以在该低频信息中提取出音频数据对应的音频特征。其中，服务器在该低频信息中提取出音频数据对应的音频特征的方式可以有多种，例如，可以在该低频信息中选取部分系数作为音频特征，比如，假设该目标频率为梅尔频率，则可以在离散余弦变换后得到的低频信息中将第 2 个到第 13 个系数作为梅尔频率倒谱系数，从而可以将每一音频帧对应的梅尔频率倒谱系数作为该音频帧对应的特征信息，根据每一音频帧的特征信息，则可以得到音频数据对应的音频特征。在一种可能的实现方式中，由于音频信号是时域连续的，分帧提取的音频帧的特征信息只反应了本帧音频的特性，为了使特征更能体现出时域连续性，服务器可以将音频数据对应的音频帧对应的特征信息在特征维度增加前后帧信息的维度，从而得到音频数据对应的音频特征。

15 在步骤 202 中，针对音频数据中包括的每个音频帧，服务器根据音频帧以及音频帧在音频数据中对应的历史音频帧，构建音频帧和历史音频帧之间的线性组合关系，基于线性组合关系中，确定音频数据对应的编码特征。

其中，服务器根据音频帧以及音频帧在音频数据中对应的历史音频帧，构建该音频帧和该历史音频帧之间的线性组合关系的方式可以有多种，例如，假设音频数据为 (Y1、Y2、Y3、Y4)，对于音频帧 Y3，可以构建该音频帧和该历史音频帧之间的线性组合关系为 $Y3=a1*Y1+a2*Y2$ ，其中，a1、a2 可以为该线性组合中的加权系数。

20 服务器在根据该音频数据的音频帧以及该音频帧在该音频数据中对应的历史音频帧，构建该音频帧和该历史音频帧之间的线性组合关系之后，便可以基于该线性组合关系中，确定该音频数据对应的编码特征。其中，服务器基于该线性组合关系中，确定该音频数据对应的编码特征的方式可以有多种，比如，服务器可以对该线性组合关系进行收敛，使得预测的音频信号与初始的音频信号误差最小，从而可以求解出加权系数，从而可以根据音频数据中每一音频帧对应的加权系数得到音频数据对应的编码特征。例如，可以采用最小均方误差等均方误差准则来对该线性组合关系进行收敛，从而可以求解出加权系数。

25 在步骤 203 中，服务器将所述音频特征以及所述编码特征进行融合，得到所述虚拟对象的融合音频特征。

其中，服务器将音频特征以及编码特征进行融合的方式可以有多种，例如，服务器可以将音频特征以及编码特征进行拼接，从而可以得到虚拟对象的融合音频特征。

在一实施例中，服务器可以采用预设特征权重对该音频特征以及该编码特征进行加权，将加权后的音频特征以及该编码特征进行拼接处理，得到该虚拟对象的融合音频特征。

30 例如，假设音频特征 M 对应的预设特征权重为 m，编码特征 N 对应的预设特征权重为 n，则可以采用预设特征权重对该音频特征以及该编码特征进行加权，得到加权后的音

频特征 $m \times M$ 以及该编码特征 $n \times N$, 从而可以将加权后的音频特征以及加权后的编码特征进行拼接处理, 得到该虚拟对象的融合音频特征 $m \times M + n \times N$ 。

在步骤 204 中, 服务器将融合音频特征进行时间间隔划分, 得到每一时间间隔的音频子特征, 按照时间间隔的顺序, 分别获取时间间隔的音频子特征对应的历史隐藏特征。

5 其中, 该时间间隔可以为表征处理融合音频特征的单位, 该音频子特征可以为以时间间隔为划分标准对融合音频特征进行划分后得到的子特征。该历史隐藏特征可以为基于历史音频子特征生成的隐藏特征, 该历史音频子特征可以为时间间隔的上一时间间隔的音频子特征。该隐藏特征可以表征当前输入的音频子特征与上一时间间隔输出的脸部数据之间的关联关系, 例如, 可以为神经网络模型中的隐藏层, 该隐藏特征可以每一时间间隔对应的
10 的隐藏层中的信息。

在步骤 205 中, 服务器根据历史隐藏特征, 对音频子特征进行特征编码, 得到音频子特征对应的语义特征, 基于语义特征, 确定融合音频特征对应的音频语义特征。

其中, 服务器根据历史隐藏特征, 对音频子特征进行特征编码的方式可以有多种, 例如, 请继续参考图 4a, 以 RNN 模型为例, 服务器可以采用 RNN 模型中的编码器结构对融合
15 音频特征进行语义特征提取。具体的, 可以参考图 4c, 其中, h_0 、 h_1 、 h_2 、 h_3 可以表示编码器结构中隐藏层中的隐藏特征, x_1 、 x_2 、 x_3 表示每一时间间隔输入的音频子特征, C 表示特征编码得到的音频语义特征。每一时间间隔对应的隐藏特征可以根据上一时间间隔的隐藏特征以及该时间间隔输入的音频子特征来确定。假设该时间间隔的输入为音频子特征 x_2 , 则音频子特征 x_2 对应的历史隐藏特征为 h_1 , 从而服务器可以采用编码器根据该历史
20 隐藏特征 h_1 , 对该音频子特征 x_2 进行特征编码, 得到该音频子特征对应的语义特征 h_2 , 也即该时间间隔对应的隐藏特征。

服务器在对该音频子特征进行特征编码之后, 便可以基于该语义特征, 确定该融合音频特征对应的音频语义特征。其中, 基于该语义特征, 确定该融合音频特征对应的音频语义特征的方式可以有多种, 例如, 服务器可以将融合音频特征的最后一个时间间隔的音频
25 子特征对应的隐藏特征(也是语义特征)作为该融合音频特征对应的音频语义特征。例如, 请继续参考图 4c, 可以将 x_3 对应的语义特征 h_3 作为音频语义特征 C 。

在步骤 206 中, 服务器获取上一时间间隔生成的历史状态数据, 根据历史状态数据以及音频语义特征, 生成当前时间间隔对应的当前状态数据, 基于当前状态数据, 确定虚拟对象的更新后状态数据。

30 其中, 服务器根据该历史状态数据以及该音频语义特征, 生成当前时间间隔对应的当前状态数据的方式可以有多种, 例如, 服务器可以在该历史状态数据中识别出目标部位参数, 基于该音频语义特征, 对该目标部位参数进行调整, 根据调整后的目标部位参数, 生成当前时间间隔对应的当前状态数据。

其中, 该目标部位参数可以为该历史状态数据中表征该虚拟人的嘴巴变化的参数。

35 在步骤 207 中, 服务器基于更新后状态数据对虚拟对象进行渲染, 得到虚拟对象的图像数据, 对更新后状态数据与音频数据进行时间对齐, 基于时间对齐后的图像数据和音频数据, 生成目标部位的状态与音频数据相匹配的视频数据。

其中,服务器基于该更新后状态数据对该虚拟对象进行渲染的方式可以有多种,例如,服务器可以采用渲染引擎(UE)以及三维图形图像软件(Blender)等渲染软件来对虚拟人的三维模型进行渲染。

在一实施例中,可以将本申请实施例提供的视频生成方法应用在广告场景中,例如,以该虚拟对象为虚拟人为例,请继续参考图 5,其中,本申请实施例提供的视频生成方法可以集成在音频驱动模块中,以此,可以上传广告的图片,并选择虚拟人形象,进而可以上传音频数据或者文本数据,在上传文本数据时,可以通过语音合成技术将文本数据转换为音频数据,在上传音频数据时,可以直接通过音频驱动模块获取更新后状态数据,并根据更新后状态数据渲染为视频数据。以此,实现了在低成本下,实现批量音频驱动 3D 虚拟人的口型参数生成,并且保证了生成的虚拟人的口型与音频数据的匹配性。

由以上可知,本申请实施例获取虚拟对象对应的初始状态数据以及音频数据,初始状态数据为虚拟对象的初始外表的数据,为了使得虚拟对象的外表数据与音频数据相匹配,可以在初始状态数据的基础上得到更新后状态数据,从而使得外表数据所表示的虚拟对象的目标部位变化与音频数据相匹配。具体地,可以从音频数据提取得到音频特征,并对音频数据进行编码得到编码特征,从而将音频特征以及编码特征进行融合,得到虚拟对象的融合音频特征。由于该融合音频特征是基于两种不同的特征(即音频特征和编码特征)进行融合得到的,音频特征和编码特征能够从不同的维度体现出音频数据的特征,故融合音频特征可以更加全面准确的表征音频数据的特征。根据这样的融合音频特征和初始状态数据,可以更加准确地得到目标部位的状态基于音频数据发生变化后得到的更新后状态数据,由此根据更新后状态数据和音频数据生成目标部位状态与音频数据相匹配的视频数据。以此,通过在音频数据中提取出音频特征和编码特征,并将音频特征和编码特征进行融合得到融合音频特征,从而根据融合音频特征和初始状态数据来得到更新后状态数据,进而根据更新后状态数据和音频数据生成目标部位状态与音频数据相匹配的视频数据,提高视频数据中目标部位与音频数据匹配的准确性,进而提高虚拟对象的视频生成准确性,进一步提升虚拟对象的视频生成效率。

为了更好地实施以上方法,本发明实施例还提供一种视频生成装置,该视频生成装置可以集成在计算机设备中,该计算机设备可以为服务器。

例如,如图 7 所示,为本申请实施例提供的视频生成装置的结构示意图,该视频生成装置可以包括获取单元 301、提取单元 302、融合单元 303、预测单元 304 和生成单元 305,如下:

获取单元 301,用于获取虚拟对象的初始状态数据以及音频数据;

提取单元 302,用于从音频数据提取得到音频特征,并对音频数据进行编码,得到编码特征;

融合单元 303,用于将该音频特征以及该编码特征进行融合,得到该虚拟对象的融合音频特征;

预测单元 304,用于根据融合音频特征和初始状态数据得到虚拟对象的更新后状态数据;

生成单元 305，用于根据该更新后状态数据和音频数据生成视频数据。

在一实施例中，该预测单元 304，包括：

语义特征提取子单元，用于对该融合音频特征进行语义特征提取，得到该音频数据对应的音频语义特征；

5 更新后状态数据生成子单元，用于根据该初始状态数据以及该音频语义特征，生成虚拟对象的更新后状态数据。

在一实施例中，该语义特征提取子单元，包括：

时间划分模块，用于将该融合音频特征进行时间间隔划分，得到每一时间间隔的音频子特征；

10 历史隐藏特征获取模块，用于按照时间间隔的顺序，分别获取每一时间间隔的音频子特征对应的历史隐藏特征，历史隐藏特征为基于历史音频子特征生成的隐藏特征，历史音频子特征为时间间隔的上一时间间隔对应的音频子特征；

特征编码模块，用于根据该历史隐藏特征，对该音频子特征进行特征编码，得到该音频子特征对应的语义特征，并基于该语义特征，确定该融合音频特征对应的音频语义特征。

15 在一实施例中，该更新后状态数据生成子单元，包括：

历史状态数据获取模块，用于获取上一时间间隔生成的历史状态数据，该历史状态数据为在该上一时间间隔，基于该初始状态数据生成的状态数据；

当前状态数据生成模块，用于根据该历史状态数据以及该音频语义特征，生成当前时间间隔对应的当前状态数据；

20 更新后状态数据确定模块，用于基于该当前状态数据，确定该更新后状态数据。

在一实施例中，该当前状态数据生成模块，包括：

识别子模块，用于在该历史状态数据中识别出目标部位参数，该目标部位参数为该历史状态数据中表征该虚拟对象的目标部位的参数；

参数调整子模块，用于参数基于该音频语义特征，对该目标部位参数进行调整；

25 当前脸部数据生成子模块，用于根据调整后的目标部位参数，生成当前时间间隔对应的当前状态数据。

在一实施例中，该生成单元 305，包括：

渲染子单元，用于基于该更新后状态数据对该虚拟对象进行渲染，得到该虚拟对象的图像数据；

30 同步子单元，用于对该更新后状态数据与该音频数据进行时间对齐；

视频生成子单元，用于基于时间对齐后的图像数据和音频数据，生成该视频数据。

在一实施例中，该提取单元 302，包括：

频率转换子单元，用于将该音频数据由当前频率转换为目标频率，并基于目标频率的音频数据，确定该音频数据对应的目标频谱；

35 倒谱分析子单元，用于根据该目标频谱对该目标频率的音频数据进行倒谱分析，得到该音频数据对应的音频特征。

在一实施例中，该频率转换子单元，包括：

分帧模块, 用于将该音频数据进行分帧处理, 得到多个音频帧;

频域转换模块, 用于针对该多个音频帧中的每个音频帧, 对该音频帧进行加窗处理, 并对加窗处理后的音频帧进行频域转换处理, 得到该音频帧的频谱信息;

滤波子单元, 用于对该频谱信息进行目标频率的滤波处理, 得到该音频数据对应的目标频谱。

在一实施例中, 该倒谱分析子单元, 包括:

非线性转换模块, 用于将该目标频谱进行非线性转换, 得到非线性转换结果;

信息提取模块, 用于对该非线性转换结果进行低频信息提取, 得到该音频数据的低频信息;

特征确定模块, 用于在该低频信息中提取出该音频数据对应的音频特征。

在一实施例中, 编码为线性预测编码, 该提取单元 302, 包括:

线性组合关系构建子单元, 用于针对该音频数据中包括的每个音频帧, 根据该音频帧以及该音频帧在该音频数据中对应的历史音频帧, 构建该音频帧和该历史音频帧之间的线性组合关系;

编码特征确定子单元, 用于基于该线性组合关系中, 确定该音频数据对应的编码特征。

在一实施例中, 该融合单元 303, 包括:

加权子单元, 用于采用预设特征权重对该音频特征以及该编码特征进行加权;

拼接子单元, 用于将加权后的音频特征以及加权后的编码特征进行拼接处理, 得到该虚拟对象的融合音频特征。

具体实施时, 以上各个单元可以作为独立的实体来实现, 也可以进行任意组合, 作为同一或若干个实体来实现, 以上各个单元的具体实施可参见前面的方法实施例, 在此不再赘述。

由以上可知, 本申请实施例通过获取单元 301 获取虚拟对象对应的初始状态数据以及音频数据; 提取单元 302 从音频数据提取得到音频特征, 并对音频数据进行编码, 得到编码特征; 融合单元 303 将音频特征以及编码特征进行融合, 得到虚拟对象的融合音频特征; 预测单元 304 根据融合音频特征和初始状态数据得到更新后状态数据; 生成单元 305 根据更新后状态数据和音频数据生成视频数据。以此, 通过在音频数据中提取出音频特征和编码特征, 并将音频特征和编码特征进行融合得到融合音频特征, 从而根据融合音频特征和初始状态数据来得到更新后状态数据, 进而根据更新后状态数据和音频数据生成目标部位状态与音频数据相匹配的视频数据, 提高视频数据中目标部位与音频数据匹配的准确性, 进而提高虚拟对象的视频生成准确性, 进一步提升虚拟对象的视频生成效率。

本申请实施例还提供一种计算机设备, 如图 8 所示, 其示出了本申请实施例所涉及的计算机设备的结构示意图, 该计算机设备可以是服务器, 具体来讲:

该计算机设备可以包括一个或者一个以上处理核心的处理器 401、一个或一个以上计算机可读存储介质的存储器 402、电源 403 和输入单元 404 等部件。本领域技术人员可以理解, 图 8 中示出的计算机设备结构并不构成对计算机设备的限定, 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。其中:

处理器 401 是该计算机设备的控制中心，利用各种接口和线路连接整个计算机设备的各个部分，通过运行或执行存储在存储器 402 内的软件程序和/或模块，以及调用存储在存储器 402 内的数据，执行计算机设备的各种功能和处理数据。可选的，处理器 401 可包括一个或多个处理核心；优选的，处理器 401 可集成应用处理器和调制解调处理器，其中，
5 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 401 中。

存储器 402 可用于存储软件程序以及模块，处理器 401 通过运行存储在存储器 402 的软件程序以及模块，从而执行各种功能应用以及虚拟对象的视频生成。存储器 402 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计算机设备的使用所创建的数据等。此外，存储器 402 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。相应地，存储器 402 还可以包括存储器控制器，以提供处理器 401 对存储器 402 的访问。
10

计算机设备还包括给各个部件供电的电源 403，优选的，电源 403 可以通过电源管理系统与处理器 401 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。电源 403 还可以包括一个或一个以上的直流或交流电源、再充电系统、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。
15

该计算机设备还可包括输入单元 404，该输入单元 404 可用于接收输入的数字或字符信息，以及产生与用户设置以及功能控制有关的键盘、鼠标、操作杆、光学或者轨迹球信号输入。
20

尽管未示出，计算机设备还可以包括显示单元等，在此不再赘述。具体在本实施例中，计算机设备中的处理器 401 会按照如下的指令，将一个或一个以上的应用程序的进程对应的可执行文件加载到存储器 402 中，并由处理器 401 来运行存储在存储器 402 中的应用程序，从而实现各种功能，如下：
25

获取虚拟对象的初始状态数据以及音频数据；
从音频数据提取得到音频特征，并对音频数据进行编码，得到编码特征；
将音频特征以及编码特征进行融合，得到虚拟对象的融合音频特征；
根据融合音频特征和初始状态数据得到虚拟对象的更新后状态数据。
以上各个操作的具体实施可参见前面的实施例，在此不作赘述。应当说明的是，本申请实施例提供的计算机设备与上文实施例中的适用于视频生成方法属于同一构思，其具体实现过程详见以上方法实施例，此处不再赘述。
30

本领域普通技术人员可以理解，上述实施例的各种方法中的全部或部分步骤可以通过指令来完成，或通过指令控制相关的硬件来完成，该指令可以存储于一计算机可读存储介质中，并由处理器进行加载和执行。

为此，本申请实施例提供一种计算机可读存储介质，其中存储有计算机程序，该计算机程序能够被处理器进行加载，以执行本申请实施例所提供的任一种虚拟对象的视频生成方法中的步骤。
35

其中，该计算机可读存储介质可以包括：只读存储器（ROM，Read Only Memory）、随机存取记忆体（RAM，Random Access Memory）、磁盘或光盘等。

5 由于该计算机可读存储介质中所存储的指令，可以执行本申请实施例所提供的任一种视频生成方法中的步骤，因此，可以实现本申请实施例所提供的任一种视频生成方法所能实现的有益效果，详见前面的实施例，在此不再赘述。

其中，根据本申请的一个方面，提供了一种计算机程序产品，该计算机程序产品包括计算机程序，该计算机程序存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机程序，处理器执行该计算机程序，使得该计算机设备执行上述实施例提供的各种可选实现方式中提供的方法。

10 以上对本申请实施例所提供的一种视频生成方法、装置和计算机可读存储介质进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

权 利 要 求

- 1、一种视频生成方法，所述方法由计算机设备执行，包括：
获取虚拟对象的初始状态数据以及音频数据；
从所述音频数据提取得到音频特征，并对所述音频数据进行编码，得到编码特征；
5 将所述音频特征以及所述编码特征进行融合，得到所述虚拟对象的融合音频特征；
根据所述融合音频特征和所述初始状态数据得到所述虚拟对象的更新后状态数据；
根据所述更新后状态数据和所述音频数据生成视频数据。
- 2、如权利要求 1 所述的视频生成方法，所述根据所述融合音频特征和所述初始状态数据得到所述虚拟对象的更新后状态数据，包括：
10 从所述融合音频特征提取得到所述音频数据对应的音频语义特征；
根据所述初始状态数据以及所述音频语义特征，生成所述更新后状态数据。
- 3、如权利要求 2 所述的视频生成方法，所述从所述融合音频特征提取得到所述音频数据对应的音频语义特征，包括：
将所述融合音频特征进行时间间隔划分，得到每一时间间隔的音频子特征；
15 按照时间间隔的顺序，分别获取所述每一时间间隔的音频子特征对应的历史隐藏特征，所述历史隐藏特征为基于历史音频子特征生成的隐藏特征，所述历史音频子特征为所述时间间隔的上一时间间隔对应的音频子特征；
根据所述历史隐藏特征，对所述音频子特征进行特征编码，得到所述音频子特征对应的语义特征，并基于所述语义特征，确定所述融合音频特征对应的音频语义特征。
- 20 4、如权利要求 2 或 3 所述的视频生成方法，所述根据所述初始状态数据以及所述音频语义特征，生成所述虚拟对象的更新后状态数据，包括：
获取上一时间间隔生成的历史脸部数据，所述历史状态数据为在所述上一时间间隔，基于所述初始状态数据生成的状态数据；
根据所述历史状态数据以及所述音频语义特征，生成当前时间间隔对应的当前状态数据；
25 据；
基于所述当前状态数据，确定所述更新后状态数据。
- 5、如权利要求 4 所述的视频生成方法，所述根据所述历史状态数据以及所述音频语义特征，生成当前时间间隔对应的当前状态数据，包括：
在所述历史状态数据中识别出目标部位参数，所述目标部位参数为所述历史状态数据
30 中表征所述虚拟对象的目标部位的参数；
基于所述音频语义特征，对所述目标部位参数进行调整；
根据调整后的目标部位参数，生成所述当前时间间隔对应的当前状态数据。
- 6、如权利要求 1-5 任一项所述的视频生成方法，所述根据所述更新后状态数据和所述音频数据生成视频数据，包括：
35 基于所述更新后状态数据对所述虚拟对象进行渲染，得到所述虚拟对象的图像数据；
对所述更新后状态数据与所述音频数据进行时间对齐；
基于时间对齐后的图像数据和所述音频数据，生成所述视频数据。

7、如权利要求 1-6 任一项所述的视频生成方法，所述从所述音频数据提取得到音频特征，包括：

将所述音频数据由当前频率转换为目标频率，并基于目标频率的音频数据，确定所述音频数据对应的目标频谱；

5 根据所述目标频谱对所述目标频率的音频数据进行倒谱分析，得到所述音频数据对应的音频特征。

8、如权利要求 7 所述的视频生成方法，所述将所述音频数据由当前频率转换为目标频率，并基于目标频率的音频数据，确定所述音频数据对应的目标频谱，包括：

将所述音频数据进行分帧处理，得到多个音频帧；

10 针对所述多个音频帧中的每个音频帧，对所述音频帧进行加窗处理，并对加窗处理后的音频帧进行频域转换处理，得到所述音频帧的频谱信息；

对所述频谱信息进行目标频率的滤波处理，得到所述音频数据对应的目标频谱。

9、如权利要求 7 或 8 所述的视频生成方法，所述根据所述目标频谱对所述目标频率的音频数据进行倒谱分析，得到所述音频数据对应的音频特征，包括：

15 将所述目标频谱进行非线性转换，得到非线性转换结果；

对所述非线性转换结果进行低频信息提取，得到所述音频数据的低频信息；

在所述低频信息中提取出所述音频数据对应的音频特征。

10、如权利要求 1-9 任一项所述的视频生成方法，所述编码为线性预测编码，所述对所述音频数据进行编码，得到编码特征，包括：

20 针对所述音频数据中包括的每个音频帧，根据所述音频帧以及所述音频帧在所述音频数据中对应的历史音频帧，构建所述音频帧和所述历史音频帧之间的线性组合关系；

基于所述线性组合关系，确定所述音频数据对应的编码特征。

11、如权利要求 1-10 任一项所述的视频生成方法，所述将所述音频特征以及所述编码特征进行融合，得到所述虚拟对象的融合音频特征，包括：

25 采用预设特征权重对所述音频特征以及所述编码特征进行加权；

将加权后的音频特征以及加权后的所述编码特征进行拼接处理，得到所述虚拟对象的融合音频特征。

12、一种视频生成装置，所述装置部署在计算机设备上，包括：

获取单元，用于获取虚拟对象的初始状态数据以及音频数据；

30 提取单元，用于从所述音频数据提取得到音频特征，并对所述音频数据进行编码，得到编码特征；

融合单元，用于将所述音频特征以及所述编码特征进行融合，得到所述虚拟对象的融合音频特征；

35 预测单元，用于根据所述融合音频特征和所述初始状态数据得到所述虚拟对象的更新后状态数据；

生成单元，用于根据所述更新后状态数据和所述音频数据生成视频数据。

13、一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序适于处理器进行加载，以执行权利要求 1 至 11 任一项所述的视频生成方法。

14、一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可以在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现权利要求 1 至 11 中任一项所述的视频生成方法。

15、一种计算机程序产品，所述计算机程序产品包括计算机程序，所述计算机程序被处理器执行时实现权利要求 1 至 11 任一项所述的视频生成方法。