(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 113689879 A
(43)申请公布日 2021.11.23

(21)申请号 202010420712.4

(22)申请日 2020.05.18

(71)申请人 北京搜狗科技发展有限公司
地址 100084 北京市海淀区中关村东路1号院9号楼搜狐网络大厦9层01房间

(72)发明人 陈伟 樊博 孟凡博

(74)专利代理机构 北京华沛德权律师事务所 11302
代理人 房德权

(51)Int.Cl.
G10L 21/10(2013.01)
G10L 25/30(2013.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

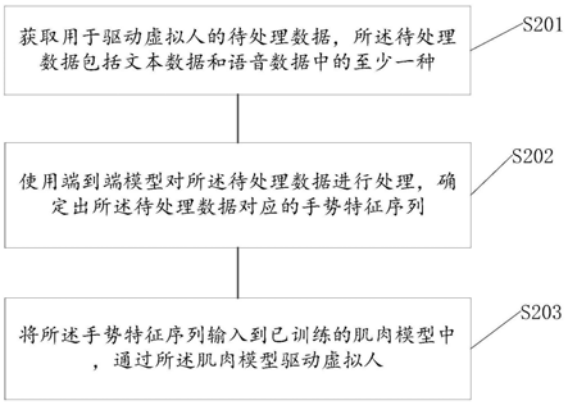
权利要求书2页 说明书14页 附图5页

(54)发明名称

实时驱动虚拟人的方法、装置、电子设备及介质

(57)摘要

本说明书实施例公开了一种实时驱动虚拟人的方法，获取用于驱动虚拟人的待处理数据，所述待处理数据包括文本数据和语音数据中的至少一种；使用端到端模型对所述待处理数据进行处理，确定出所述待处理数据对应的手势特征序列；将所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人；由于端到端模型能够在更短时间内获取手势特征序列；再将手势特征序列输入到肌肉模型中，直接驱动虚拟人，极大的降低了其计算量和数据传输量，且还提高了计算效率，使得驱动虚拟人的实时性得到极大的提高，从而能够实现实时驱动虚拟人进行手语输出。



CN 113689879 A

1.一种实时驱动虚拟人的方法,其特征在于,包括:

获取用于驱动虚拟人的待处理数据,所述待处理数据包括文本数据和语音数据中的至少一种;

使用端到端模型对所述待处理数据进行处理,确定出所述待处理数据对应的手势特征序列;

将所述手势特征序列输入到已训练的肌肉模型中,通过所述肌肉模型驱动虚拟人;

其中,所述使用端到端模型对所述待处理数据进行处理,包括:

获取所述待处理数据的文本特征和时长特征;

根据所述文本特征和所述时长特征,确定出所述手势特征序列。

2.如权利要求1所述的方法,其特征在于,所述获取所述待处理数据的文本特征和时长特征,包括:

通过fastspeech模型获取所述文本特征;

通过时长模型获取所述时长特征,其中,所述时长模型为深度学习模型。

3.如权利要求2所述的方法,其特征在于,若所述fastspeech模型输出面部特征序列和手势特征序列,所述根据所述文本特征和所述时长特征,确定出所述声学特征序列,包括:

将所述文本特征和所述时长特征输入到所述fastspeech模型中,得到所述面部特征序列和所述手势特征序列。

4.如权利要求3所述的方法,其特征在于,所述将所述手势特征序列输入到已训练的肌肉模型中,包括:

将所述面部特征序列和所述手势特征序列进行融合,得到融合特征序列;

将所述融合特征序列输入到所述肌肉模型中。

5.如权利要求4所述的方法,其特征在于,所述将所述面部特征序列和所述手势特征序列进行融合,得到融合特征序列,包括:

基于所述时长特征,将所述面部特征序列和所述手势特征序列进行融合,得到所述融合特征序列。

6.如权利要求5所述的方法,其特征在于,所述面部特征序列对应的面部特征包括表情特征和唇部特征。

7.一种实时驱动虚拟人的装置,其特征在于,包括:

数据获取模块,用于获取用于驱动虚拟人的待处理数据,所述待处理数据包括文本数据和语音数据中的至少一种;

数据处理模块,用于使用端到端模型对所述待处理数据进行处理,确定出所述待处理数据对应的手势特征序列;

虚拟人驱动模块,用于将所述手势特征序列输入到已训练的肌肉模型中,通过所述肌肉模型驱动虚拟人;

其中,所述数据处理模块,用于获取所述待处理数据的文本特征和时长特征;根根据所述文本特征和所述时长特征,确定出所述手势特征序列。

8.如权利要求7所述的装置,其特征在于,所述数据处理模块,用于通过fastspeech模型获取所述文本特征;通过时长模型获取所述时长特征,其中,所述时长模型为深度学习模型。

9.一种用于数据处理的装置,其特征在于,包括有存储器,以及一个或者一个以上的程序,其中一个或者一个以上程序存储于存储器中,且经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含如权利要求1-6任一权项所述的方法步骤。

10.一种机器可读介质,其上存储有指令,当由一个或多个处理器执行时,使得装置执行如权利要求1至6中一个或多个所述的实时驱动虚拟人的方法。

实时驱动虚拟人的方法、装置、电子设备及介质

技术领域

[0001] 本说明书实施例涉及虚拟人处理技术领域，尤其涉及一种实时驱动虚拟人的方法、装置、电子设备及介质。

背景技术

[0002] 数字人类(Digital Human)简称数字人，是利用计算机模拟真实人类的一种综合性的渲染技术，也被称为虚拟人类、超写实人类、照片级人类。由于人对真人太熟悉了，通过花费大量时间可以获取使得3D静态模型很真，但在驱动3D静态模型进行动作时，即使是一个细微的表情都会重新建模，由于模型的真实度非常高会导致建模会需要进行大量的数据进行计算，其计算过程较长，通常模型的一个动作可能需要一个小时或几个小时的计算才能实现，导致驱动的实时性能非常差。

发明内容

[0003] 本说明书实施例提供了一种实时驱动虚拟人的方法、装置、电子设备及介质，使得驱动虚拟人的实时性提高。

[0004] 本说明书实施例第一方面提供了一种实时驱动虚拟人的方法，包括：

[0005] 获取用于驱动虚拟人的待处理数据，所述待处理数据包括文本数据和语音数据中的至少一种；

[0006] 使用端到端模型对所述待处理数据进行处理，确定出所述待处理数据对应的手势特征序列；

[0007] 将所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人；

[0008] 其中，所述使用端到端模型对所述待处理数据进行处理，包括：

[0009] 获取所述待处理数据的文本特征和时长特征；

[0010] 根据所述文本特征和所述时长特征，确定出所述手势特征序列。

[0011] 可选的，所述获取所述待处理数据的文本特征和时长特征，包括：

[0012] 通过fastspeech模型获取所述文本特征；

[0013] 通过时长模型获取所述时长特征，其中，所述时长模型为深度学习模型。

[0014] 可选的，若所述fastspeech模型输出面部特征序列和手势特征序列，所述根据所述文本特征和所述时长特征，确定出所述声学特征序列，包括：

[0015] 将所述文本特征和所述时长特征输入到所述fastspeech模型中，得到所述面部特征序列和所述手势特征序列。

[0016] 可选的，所述将所述手势特征序列输入到已训练的肌肉模型中，包括：

[0017] 将所述面部特征序列和所述手势特征序列进行融合，得到融合特征序列；

[0018] 将所述融合特征序列输入到所述肌肉模型中。

[0019] 可选的，所述将所述面部特征序列和所述手势特征序列进行融合，得到融合特征

序列,包括:

[0020] 基于所述时长特征,将所述面部特征序列和所述手势特征序列进行融合,得到所述融合特征序列。

[0021] 可选的,所述面部特征序列对应的面部特征包括表情特征和唇部特征。

[0022] 本说明书实施例第二方面提供了一种实时驱动虚拟人的装置,包括:

[0023] 数据获取模块,用于获取用于驱动虚拟人的待处理数据,所述待处理数据包括文本数据和语音数据中的至少一种;

[0024] 数据处理模块,用于使用端到端模型对所述待处理数据进行处理,确定出所述待处理数据对应的手势特征序列;

[0025] 虚拟人驱动模块,用于将所述手势特征序列输入到已训练的肌肉模型中,通过所述肌肉模型驱动虚拟人;

[0026] 其中,所述数据处理模块,用于获取所述待处理数据的文本特征和时长特征;根根据所述文本特征和所述时长特征,确定出所述手势特征序列。

[0027] 可选的,所述数据处理模块,用于通过fastspeech模型获取所述文本特征;通过时长模型获取所述时长特征,其中,所述时长模型为深度学习模型。

[0028] 可选的,所述数据处理模块,若所述fastspeech模型输出面部特征序列和手势特征序列,用于将所述文本特征和所述时长特征输入到所述fastspeech模型中,得到所述面部特征序列和所述手势特征序列。

[0029] 可选的,所述虚拟人驱动模块,用于将所述面部特征序列和所述手势特征序列进行融合,得到融合特征序列;将所述融合特征序列输入到所述肌肉模型中。

[0030] 可选的,所述虚拟人驱动模块,用于基于所述时长特征,将所述面部特征序列和所述手势特征序列进行融合,得到所述融合特征序列。

[0031] 可选的,所述面部特征序列对应的面部特征包括表情特征和唇部特征。

[0032] 本说明书实施例第三方面提供了一种用于数据处理的装置,其特征在于,包括有存储器,以及一个或者一个以上的程序,其中一个或者一个以上程序存储于存储器中,且经配置以由一个或者一个以上处理器执行所述一个或者一个以上程序包含用于如上述实时驱动虚拟人的方法步骤。

[0033] 本说明书实施例第四方面提供了一种机器可读介质,其上存储有指令,当由一个或多个处理器执行时,使得装置执行如上述实时驱动虚拟人的方法。

[0034] 本说明书实施例的有益效果如下:

[0035] 基于上述技术方案,在获取待处理数据之后,使用端到端模型对待处理数据进行处理,得到手势特征序列;再将所述手势特征序列输入到已训练的肌肉模型中,通过所述肌肉模型驱动虚拟人;由于端到端模型输入的是待处理数据的原始数据,而直接输出手势特征序列,其能够更好的利用和适应新的硬件(比如GPU)并行计算能力,运算速度更快;即,能够在更短时间内获取手势特征序列;再将手势特征序列输入到肌肉模型中,直接驱动虚拟人,是在创建虚拟人之后,直接通过声学特征序列来控制虚拟人进行语音输出,并同时通过面部特征序列和手势特征序列控制虚拟人的手势动作,与需要重新对虚拟人建模相比,极大的降低了其计算量和数据传输量,且还提高了计算效率,使得驱动虚拟人的实时性得到极大的提高,从而能够实现实时驱动虚拟人进行手语输出。

附图说明

[0036] 图1为本说明书实施例中对输出声学特征序列的端到端模型进行训练的训练流程图；

[0037] 图2为本说明书实施例中实时驱动虚拟人的方法的第一种流程图；

[0038] 图3为本说明书实施例中第一fastspeech模型输出声学特征序列的步骤流程图；

[0039] 图4为本说明书实施例中实时驱动虚拟人的方法的第二种流程图；

[0040] 图5为本说明书实施例中实时驱动虚拟人的装置的结构示意图；

[0041] 图6为本说明书实施例中用于实时驱动虚拟人的装置作为设备时的结构框图；

[0042] 图7为本说明书实施例中一些实施例中服务端的结构框图。

具体实施方式

[0043] 为了更好的理解上述技术方案，下面通过附图以及具体实施例对本说明书实施例的技术方案做详细的说明，应当理解本说明书实施例以及实施例中的具体特征是对本说明书实施例技术方案的详细的说明，而不是对本说明书技术方案的限定，在不冲突的情况下，本说明书实施例以及实施例中的技术特征可以相互组合。

[0044] 针对虚拟人在驱动时需要耗费大量时间的技术问题，本发明实施例提供了一种实时驱动虚拟人的方案，该方案用于实时驱动虚拟人，具体可以包括：获取用于驱动虚拟人的待处理数据，所述待处理数据包括文本数据和语音数据中的至少一种；使用端到端模型对所述待处理数据进行处理，确定出所述待处理数据对应的手势特征序列；将所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人；

[0045] 其中，所述使用端到端模型对所述待处理数据进行处理，确定出所述待处理数据对应的手势特征序列，包括：获取所述待处理数据的文本特征和时长特征；根据所述文本特征和所述时长特征，确定出所述手势特征序列。

[0046] 本发明实施例中的虚拟人具体可以是高仿真虚拟人，与真人的差异较小；本发明人实施例中虚拟人可以应用于新闻播报场景、教学场景、医疗场景、客服场景、法律场景和会议场景等内容表达场景。

[0047] 本发明实施例中待处理数据可以是文本数据，也可以是语音数据，也可以是文本数据和语义数据同时存在，本说明书不作具体限制。

[0048] 例如，在新闻播报场景，需要获取驱动虚拟人的待播报的新闻稿，此时，新闻稿为待处理数据，且新闻稿可以是由人工或机器编辑的文本，以及在人工或机器编辑文本之后，获取编辑的文本作为新闻稿，其中，新闻稿进行手势播报。

[0049] 本发明实施例中，在使用端到端模型对所述待处理数据进行处理之前，还需通过样本对端到端模型进行训练，得到已训练的端到端模型；在得到已训练的端到端模型之后，再使用已训练的端到端模型对所述待处理数据进行处理。

[0050] 本发明实施例中端到端模型包括两种训练方法，其中一种训练方法训练出的端到端模型输出的声学特征序列，另一种训练方法训练出的端到端模型输出的手势特征序列；以及端到端模型具体可以为fastspeech模型。

[0051] 其中，在对输出声学特征序列的端到端模型进行训练时，其训练样本可以是文本和语音数据，还可以视频数据；针对训练样本集中每个训练样本，其训练步骤具体如图1所

示，首先执行步骤A1，获取训练样本的声学特征101和文本特征102，其中，文本特征101可以为音素级别。具体地，可以将训练样本的特征数据映射到端到端模型中的嵌入(embedding)层中，得到声学特征101和文本特征102；然后执行步骤A2，通过前馈变压器103(Feed Forward Transformer)处理声学特征101和文本特征102，得到声学向量104和文本编码特征105，其中，声学向量104可以是句子的声学向量，也可以词的声学向量，文本编码特征105同样是音素级别；接下执行步骤A3，将声学向量104和文本编码特征105进行对齐，得到对齐后的文本编码特征106，可以使用持续时间预测器将声学向量104和文本编码特征105进行对齐，其中，文本编码特征105具体为音素特征，声学向量104可以是梅尔频谱图，如此，可以使用持续时间预测器将因素特征和梅尔频谱图进行对齐；接下来执行步骤A4，对对齐后的文本编码特征106进行解码107，获取声学特征序列108，此时，可以使用长度调节器通过延长或缩短音素持续时间来轻松确定语音速度，从而确定生成的梅尔频谱图的长度，还可以通过在相邻音素之间添加间隔来控制部分韵律；根据确定出的梅尔频谱图的长度和音素间隔时间，获取到声学特征序列。

[0052] 在对输出声学特征序列的端到端模型进行训练时，其训练样本集例如可以包含13,100个语音频剪辑和相应的文本记录，音频总长度约为24小时。此时，将训练样本集随机分为3组：用于训练的12500个样本，用于验证的300个样本和用于测试的300个样本。为了减轻发音错误的问题，使用音素转换工具将文本序列转换为音素序列；对于语音数据，将原始波形转换为梅尔频谱图；然后使用12500个样本对端到端模型进行训练进行训练，在训练完成之后，使用300个验证样本对训练得到的端到端模型进行验证；在验证符合验证要求之后，使用300个测试样本对端到端模型进行测试，若测试符合测试条件，则得到已训练的端到端模型。

[0053] 若对端到端模型进行验证未符合验证要求，则使用训练样本再次对端到端模型训练，直至训练后的端到端模型符合验证要求；并对验证符合要求的端到端模型进行测试，直至训练后的端到端模型既符合验证要求也符合测试条件，则将训练后的端到端模型作为最终的模型，即为已训练的端到端模型。

[0054] 以及，在对输出手势特征序列的端到端模型进行训练时，其训练样本可以是真人视频数据和真人动作数据；针对训练样本集中每个训练样本，其训练步骤具体包括，首先执行步骤B1，获取训练样本的手势特征和文本特征，其中，文本特征可以为音素级别。具体地，可以将训练样本的特征数据映射到端到端模型中的嵌入(embedding)层中，得到手势特征和文本特征；然后执行步骤B2，通过前馈变压器(Feed Forward Transformer)处理手势特征和文本特征，得到手势特征向量和文本编码特征，其中，手势特征向量可以是肌肉动作向量，文本编码特征同样是音素级别；接下执行步骤B3，将手势特征向量与文本编码特征进行对齐，可以使用持续时间预测器将手势特征向量与文本编码特征进行对齐，其中，文本编码特征具体为音素特征；接下来执行步骤B4，获取手势特征序列，此时，可以使用长度调节器通过延长或缩短音素持续时间来对齐手势动作，从而得到手势特征序列。

[0055] 本发明实施例中文本特征可以包括：音素特征、和/或、语义特征等。进一步的，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，一个动作构成一个音素。音素可以包括：元音与辅音。可选地，特定的音素特征对应特定的唇部特征、表情特征和手势特征等。

[0056] 以及，语义是待处理文本所对应的现实世界中的事物所代表的概念的含义，以及这些含义之间的关系，是待处理文本在某个领域上的解释和逻辑表示。可选地，特定的语义特征对应特定的手势特征等。

[0057] 在对输出手势特征序列的端到端模型进行训练时，其训练样本集包括的真人动作数据或者真人视频数据，其训练过程参考对输出声学特征序列的端到端模型进行训练的训练过程，为了说明书的简洁，在此就不再赘述了。

[0058] 如此，在得到待处理数据之后，可以使用第二端到端模型的嵌入层获取所述待处理数据的文本特征，再获取所述待处理数据的时长特征，将所述文本特征和所述时长特征输入到第二端到端模型中，得到手势特征序列。

[0059] 当然，还可以在得到待处理数据之后，先利用第一端到端模型的嵌入层获取所述待处理数据的文本特征，再获取所述待处理数据的时长特征，将所述文本特征和所述时长特征输入到第一端到端模型中，得到所述声学特征序列；相应地，还可以同时或之后利用第二端到端模型的嵌入层获取所述待处理数据的文本特征，再获取所述待处理数据的时长特征，将所述文本特征和所述时长特征输入到第二端到端模型中，得到手势特征序列；当然，也可以直接利用前面获取的文本特征和时长特征直接输入到第二端到端模型中，得到手势特征序列。本说明书实施例中，第一端到端模型和第二端到端模型可以同时处理数据，也可以是第一端到端模型先处理数据，还可以是第二端到端模型先处理数据，本说明书不作具体限制。

[0060] 本发明实施例中，时长特征可用于表征文本所对应音素的时长。时长特征能够刻画出语音中的抑扬顿挫与轻重缓急，进而可以提高合成语音的表现力和自然度。可选地，可以利用时长模型，确定待处理数据对应的时长特征。时长模型的输入可以为：带有重音标注的音素特征，输出为音素时长。时长模型可以为对带有时长信息的语音样本进行学习得到，例如，可以是卷积神经网络(Convolutional Neural Networks，以下简称CNN)和深度神经网络(Deep Neural Networks，以下简称DNN)等深度学习模型，本发明实施例对于具体的时长模型不加以限制。

[0061] 以及，在获取到所述手势特征序列之后，将得到所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人，以驱动虚拟人通过手势动作来表达所述待处理数据的语义，即将所述待处理数据通过手语进行输出。

[0062] 手势特征是指用手部位的协调活动来传达人物的思想，形象地借以表情达意。

[0063] 本发明实施例中，在使用已训练的肌肉模型之前，还需进行模型训练，得到已训练的肌肉模型；在得到已训练的肌肉模型之后，再使用已训练的肌肉模型对所述待处理数据的文本特征进行处理。

[0064] 本发明实施例中已训练的肌肉模型在进行模型训练时，首先根据人的手指肌肉来创建肌肉模型，在获取其训练样本，其训练样本可以是真人视频数据和真人动作数据；针对训练样本集中每个训练样本，其训练步骤包括：

[0065] 首先执行步骤C1，获取每个训练样本的手指肌肉特征；然后执行步骤C2，使用每个训练样本的手指肌肉特征对肌肉模型进行训练；以及，在训练完成之后，执行步骤C3，使用验证样本对训练得到的肌肉模型进行验证；在验证符合验证要求之后，再使用测试样本对训练得到的肌肉模型进行测试，若测试符合测试条件，则得到已训练的肌肉模型。

[0066] 若对训练得到的肌肉模型进行验证未符合验证要求，则使用训练样本再次对肌肉模型训练，直至训练后的肌肉模型符合验证要求；并对验证符合要求的肌肉模型进行测试，直至训练后的肌肉模型既符合验证要求也符合测试条件，则将训练后的肌肉模型作为最终的模型，即为已训练的肌肉模型。

[0067] 以及，在创建肌肉模型时，以手指肌肉特征为例，使用多边形网络进行近似抽象的肌肉控制，可以使用两类肌肉，一种线性肌肉，用于拉伸；一种括约肌，用于挤压；两种肌肉只在一点与网格空间相联系，有方向指定(两种肌肉变形时都是计算某一点的角位移和径向位移)，因此肌肉的控制独立于具体的面部拓扑，使得面部表情能够更逼真且更细腻；相应地，手指肌肉也使用多边形网络进行近似抽象的肌肉控制，从而能够确保手势动作更准确。

[0068] 由于端到端模型的前馈变压器采用的自注意力机制是一种通过其上下文来理解当前词的创新方法，语义特征的提取能力更强。在实际应用中，这个特性意味着对于句子中的同音字或词，新的算法能根据它周围的词和前后的句子来判断究竟应该是哪个(比如洗澡和洗枣)，从而得到更准确的结果；而且端到端模型解决了传统的语音识别方案中各部分任务独立，无法联合优化的问题。单一神经网络的框架变得更简单，随着模型层数更深，训练数据越大，准确率越高；第三，端到端模型采用新的神经网络结构，其可以更好地利用和适应新的硬件(比如GPU)并行计算能力，运算速度更快。这意味着转写同样时长的语音，基于新网络结构的算法模型可以在更短的时间内完成，也更能满足实时转写的需求。

[0069] 本发明实施例在获取待处理数据之后，使用端到端模型对待处理数据进行处理，得到手势特征序列；再将所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人；由于端到端模型输入的是待处理数据的原始数据，而直接输出手势特征序列，其能够更好的利用和适应新的硬件(比如GPU)并行计算能力，运算速度更快；即，能够在更短时间内获取手势特征序列；再将手势特征序列输入到肌肉模型中，直接驱动虚拟人，是在创建虚拟人之后，直接通过声学特征序列来控制虚拟人进行语音输出，并同时通过面部特征序列和手势特征序列控制虚拟人的面部表情和手势动作，与需要重新对虚拟人建模相比，极大的降低了其计算量和数据传输量，且还提高了计算效率，使得驱动虚拟人的实时性得到极大的提高，从而能够实现实时驱动虚拟人进行手语输出。

[0070] 而且，由于采用端到端模型来获取手势特征序列时，使用了时长特征，而时长特征能够提高声学特征序列和手势特征序列之间的同步性，从而在同步性的提高的基础上，使用手势特征序列来驱动虚拟人时，能够使得虚拟人的声音输出与手势特征匹配的精确度更高。

[0071] 方法实施例一

[0072] 参照图2，示出了本发明的一种实时驱动虚拟人的方法实施例一的步骤流程图，具体可以包括如下步骤：

[0073] S201、获取用于驱动虚拟人的待处理数据，所述待处理数据包括文本数据和语音数据中的至少一种；

[0074] S202、使用端到端模型对所述待处理数据进行处理，确定出所述待处理数据对应的手势特征序列；

[0075] S203、将所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动

虚拟人；

[0076] 其中，步骤S201包括：

[0077] 步骤S2011、获取所述待处理数据的文本特征和时长特征；

[0078] 步骤S2012、根据所述文本特征和所述时长特征，确定出所述手势特征序列。

[0079] 步骤S201中，对于客户端而言，可以接收用户上传的待处理数据；对于服务端而言，可以接收客户端发送的待处理数据。可以理解，任意的第一设备可以从第二设备接收待处理文本，本发明实施例对于待处理数据的具体传输方式不加以限制。

[0080] 若待处理数据为文本数据，则直接使用步骤S202对待处理数据进行处理；若待处理数据为语音数据，则将待处理数据转换成文本数据之后，使用步骤S202对转换后的文本数据进行处理。

[0081] 步骤S202中，首先需要训练出端到端模型，其中，端到端模型包括两种训练方法，其中一种训练方法训练出的端到端模型输出的声学特征序列，另一种训练方法训练出的端到端模型输出的手势特征序列；以及端到端模型具体可以为fastspeech模型。

[0082] 以及训练出输出声学特征序列的端到端模型作为第一端到端模型，其训练过程中具体参考上述步骤A1-A4的叙述；训练出输出手势特征序列的端到端模型作为第二端到端模型，其训练过程参考步骤B1-B4的叙述。

[0083] 若端到端模型为fastspeech模型，则训练得到第一fastspeech模型和第二fastspeech模型之后，使用任意一个fastspeech模型获取到待处理数据的文本特征；再使用时长模型获取到时长特征，其中，时长模型可以是CNN和DNN等深度学习模型。

[0084] 具体地，如图3所示，以第一fastspeech模型获取手势特征序列为例，其步骤包括：通过第一fastspeech模型的嵌入层获取待处理数据的文本特征301，通过前馈变压器302对文本特征301进行编码，得到文本编码特征303；此时，通过时长模型304对文本编码特征303处理，得到时长特征305，其中，时长特征304可用于表征文本编码特征30中每个音素的时长；然后通过时长特征305对文本编码特征303进行对齐，得到对齐后的文本编码特征306；对对齐后的文本编码特征306进行解码307并预测，得到声音特征序列307。

[0085] 其中，文本编码特征303是音素级别，对齐后的文本编码特征306可以是帧级，也可以是音素级别，本发明实施例不作具体限制。

[0086] 相应地，使用第二fastspeech模型获取手势特征序列过程中，可以通过第二fastspeech模型的嵌入层获取待处理数据的文本特征；再通过前馈变压器对文本特征进行编码，得到文本编码特征；此时，通过时长模型对文本编码特征处理，得到时长特征，其中，时长特征对文本编码特征进行对齐，得到对齐后的文本编码特征；对对齐后的文本编码特征进行解码后进行手势预测，得到手势特征序列。

[0087] 以及，还可以使用以第一fastspeech模型获取声学特征序列和使用时长模型获取时长特征，通过时长特征将所述声学特征序列和手势特征序列对齐，使得在通过声学特征序列和手势特征序列输入肌肉模型中驱动虚拟人时，虚拟人的声音播报和手语播报保持同步。

[0088] 接下来执行步骤S203，将所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人，以驱动虚拟人通过手语来输出所述待处理数据。

[0089] 具体来讲，根据所述时长特征，将所述声学特征序列和手势特征序列对齐，使得在

通过声学特征序列和手势特征序列输入肌肉模型中驱动虚拟人时，虚拟人的声音播报和手语播报保持同步。

[0090] 例如，在声学特征序列在说“再见”时，虚拟人的手语输出的“再见手语”，从而保持声音和手语的一致性；相应地，在声音特征序列在说“春天百花开放”，虚拟人的手语输出的“春天百花开放”，从而保持声音和手语的一致性。

[0091] 本发明实施例在获取待处理数据之后，使用端到端模型对待处理数据进行处理，得到手势特征序列；再将所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人；由于端到端模型输入的是待处理数据的原始数据，而直接输出手势特征序列，其能够更好的利用和适应新的硬件(比如GPU)并行计算能力，运算速度更快；即，能够在更短时间内获取手势特征序列；再将手势特征序列输入到肌肉模型中，直接驱动虚拟人，是在创建虚拟人之后，直接通过声学特征序列来控制虚拟人进行语音输出，并同时通过面部特征序列和手势特征序列控制虚拟人的手势动作，与需要重新对虚拟人建模相比，极大的降低了其计算量和数据传输量，且还提高了计算效率，使得驱动虚拟人的实时性得到极大的提高，从而能够实现实时驱动虚拟人进行手语输出。

[0092] 而且，由于采用端到端模型来获取手势特征序列时，使用了时长特征，而时长特征能够提高声学特征序列和手势特征序列之间的同步性，从而在同步性的提高的基础上，使用手势特征序列来驱动虚拟人时，能够使得声音输出与虚拟人的手语输出匹配的精确度更高。

[0093] 方法实施例二

[0094] 参照图4，示出了本发明的一种实时驱动虚拟人的方法实施例一的步骤流程图，具体可以包括如下步骤：

[0095] S401、获取用于驱动虚拟人的待处理数据，所述待处理数据包括文本数据和语音数据中的至少一种；

[0096] S402、使用端到端模型对所述待处理数据进行处理，确定出所述待处理数据对应的面部特征序列和手势特征序列；

[0097] S403、将所述面部特征序列和所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人；

[0098] 其中，步骤S401包括：

[0099] 步骤S4011、获取所述待处理数据的文本特征和时长特征；

[0100] 步骤S4012、根据所述文本特征和所述时长特征，确定出所述面部特征序列和所述手势特征序列。

[0101] 步骤S401中，对于客户端而言，可以接收用户上传的待处理数据；对于服务端而言，可以接收客户端发送的待处理数据。可以理解，任意的第一设备可以从第二设备接收待处理文本，本发明实施例对于待处理数据的具体传输方式不加以限制。

[0102] 若待处理数据为文本数据，则直接使用步骤S402对待处理数据进行处理；若待处理数据为语音数据，则将待处理数据转换成文本数据之后，使用步骤S402对转换后的文本数据进行处理。

[0103] 步骤S402中，首先需要训练出输出面部特征序列和手势特征序列模型，此时，在对输出面部特征序列和手势特征序列的端到端模型进行训练时，其训练样本可以是真人视频

数据和真人动作数据;针对训练样本集中每个训练样本,其训练步骤具体包括,首先执行步骤D1,获取训练样本的面部特征、手势特征和文本特征,其中,文本特征可以为音素级别。具体地,可以将训练样本的特征数据映射到端到端模型中的嵌入(embedding)层中,得到面部特征、手势特征和文本特征;然后执行步骤D2,通过前馈变压器(Feed Forward Transformer)处理面部特征、手势特征和文本特征,得到面部特征向量、手势特征向量和文本编码特征,其中,面部特征向量是用于进行面部表情的特征表示,手势特征向量可以是肌肉动作向量,文本编码特征同样是音素级别;接下执行步骤D3,将面部特征向量和手势特征向量,与文本编码特征进行对齐,可以使用持续时间预测器将面部特征向量和手势特征向量,与文本编码特征进行对齐,其中,文本编码特征具体为音素特征;接下来执行步骤D4,获取面部特征序列和手势特征序列,此时,可以使用长度调节器通过延长或缩短音素持续时间来对齐面部表情和手势动作,从而得到面部特征序列和手势特征序列。

[0104] 本发明实施例中文本特征可以包括:音素特征、和/或、语义特征等。进一步的,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,一个动作构成一个音素。音素可以包括:元音与辅音。可选地,特定的音素特征对应特定的唇部特征、表情特征和手势特征等。

[0105] 以及,语义是待处理文本所对应的现实世界中的事物所代表的概念的含义,以及这些含义之间的关系,是待处理文本在某个领域上的解释和逻辑表示。可选地,特定的语义特征对应特定的手势特征等。

[0106] 在对输出面部特征序列和手势特征序列的端到端模型进行训练时,其训练样本集包括的真人动作数据或者真人视频数据,其训练过程参考对输出声学特征序列的端到端模型进行训练的训练过程,为了说明书的简洁,在此就不再赘述了。

[0107] 以及,在训练得到输出面部特征序列和手势特征序列的端到端模型之后,将得到的输出面部特征序列和手势特征序列的端到端模型作为第三端到端模型。

[0108] 如此,在得到待处理数据之后,可以使用第三端到端模型的嵌入层获取所述待处理数据的文本特征,再获取所述待处理数据的时长特征,将所述文本特征和所述时长特征输入到第三端到端模型中,得到面部特征序列和手势特征序列。

[0109] 以及,在得到待处理数据之后,可以使用第三端到端模型的嵌入层获取所述待处理数据的文本特征,再获取所述待处理数据的时长特征,将所述文本特征和所述时长特征输入到第三端到端模型中,得到面部特征序列和手势特征序列。

[0110] 当然,还可以在得到待处理数据之后,先利用第一端到端模型的嵌入层获取所述待处理数据的文本特征,再获取所述待处理数据的时长特征,将所述文本特征和所述时长特征输入到第一端到端模型中,得到所述声学特征序列;相应地,还可以同时或之后利用第三端到端模型的嵌入层获取所述待处理数据的文本特征,再获取所述待处理数据的时长特征,将所述文本特征和所述时长特征输入到第二端到端模型中,得到面部特征序列和手势特征序列;当然,也可以直接利用前面获取的文本特征和时长特征直接输入到第三端到端模型中,得到面部特征序列和手势特征序列。本说明书实施例中,第一端到端模型和第三端到端模型可以同时处理数据,也可以是第一端到端模型先处理数据,还可以是第三端到端模型先处理数据,本说明书不作具体限制。

[0111] 本发明实施例中,时长特征可用于表征文本所对应音素的时长。时长特征能够刻

画出语音中的抑扬顿挫与轻重缓急，进而可以提高合成语音的表现力和自然度。可选地，可以利用时长模型，确定待处理数据对应的时长特征。时长模型的输入可以为：带有重音标注的音素特征，输出为音素时长。时长模型可以为对带有时长信息的语音样本进行学习得到，例如，可以是卷积神经网络(Convolutional Neural Networks，以下简称CNN)和深度神经网络(Deep Neural Networks，以下简称DNN)等深度学习模型，本发明实施例对于具体的时长模型不加以限制。

[0112] 以及，在获取到所述面部特征序列和所述手势特征序列之后，将得到所述面部特征序列和所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人，以驱动虚拟人通过手势动作来表达所述待处理数据的语义，即将所述待处理数据通过手语进行输出，并根据所述待处理数据的语义虚拟人会呈现不同表情特征。

[0113] 本发明实施例中，面部特征包括表情特征和唇部特征，其中，表情，表达感情、情意，可以指表现在面部的思想感情。表情特征通常是针对整个面部的。唇部特征可以专门针对唇部，而且跟文本的文本内容、语音、发音方式等都有关系，从而可以通过面部特征能够促使面部表情更逼真且更细腻。

[0114] 相应地，还可以使用以第一fastspeech模型获取声学特征序列和使用时长模型获取时长特征，通过时长特征将所述声学特征序列，与面部特征序列和手势特征序列对齐，使得在通过声学特征序列，面部特征序列和手势特征序列输入肌肉模型中驱动虚拟人时，虚拟人的声音播报与面部表情和手语播报保持同步。

[0115] 相应地，使用第三fastspeech模型获取手势特征序列过程中，可以通过第三fastspeech模型的嵌入层获取待处理数据的文本特征；再通过前馈变压器对文本特征进行编码，得到文本编码特征；此时，通过时长模型对文本编码特征处理，得到时长特征，其中，时长特征对文本编码特征进行对齐，得到对齐后的文本编码特征；对对齐后的文本编码特征进行解码后进行面部预测和手势预测，得到面部特征序列和手势特征序列。

[0116] 以及，还可以使用以第一fastspeech模型获取声学特征序列和使用时长模型获取时长特征，通过时长特征将声学特征序列与，面部特征序列和手势特征序列对齐，使得在通过声学特征序列，面部特征序列和手势特征序列输入肌肉模型中驱动虚拟人时，虚拟人的声音播报，面部表情和手语播报保持同步。

[0117] 接下来执行步骤S203，将面部特征序列和手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人，以驱动虚拟人同时控制面部表情并输出手语。

[0118] 具体来讲，根据所述时长特征，将所述声学特征序列和手势特征序列对齐，使得在通过声学特征序列和手势特征序列输入肌肉模型中驱动虚拟人时，虚拟人的声音播报和手语播报保持同步。

[0119] 例如，在声学特征序列在说“再见”时，虚拟人的手语输出的“再见手语”并面部呈现微笑，从而保持声音，与面部表情和手语的一致性；相应地，在声音特征序列在说“某人受伤”，虚拟人的手语输出的“某人受伤”并面部呈现悲伤，从而保持声音，与面部表情和手语的一致性。

[0120] 本发明实施例在获取待处理数据之后，使用端到端模型对待处理数据进行处理，得到面部特征序列和手势特征序列；再将所述面部特征序列和手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人；由于端到端模型输入的是待处理数据的原

始数据，而直接输出面部特征序列和手势特征序列，其能够更好的利用和适应新的硬件(比如GPU)并行计算能力，运算速度更快；即，能够在更短时间内获取面部特征序列和手势特征序列；再将面部特征序列和手势特征序列输入到肌肉模型中，直接驱动虚拟人，是在创建虚拟人之后，在通过声学特征序列来控制虚拟人进行语音输出的同时通过面部特征序列和手势特征序列控制虚拟人的面部表情和手势动作，与需要重新对虚拟人建模相比，极大的降低了其计算量和数据传输量，且还提高了计算效率，从而能实现了实时驱动虚拟人以手语输出。

[0121] 而且，由于采用端到端模型来获取面部特征序列和手势特征序列时，使用了时长特征，而时长特征能够提高声学特征序列，与面部特征序列和手势特征序列之间的同步性，从而在同步性的提高的基础上，使用面部特征序列和和手势特征序列来驱动虚拟人时，能够使得声音输出与面部表情和手语匹配的精确度更高。

[0122] 装置实施例

[0123] 参照图5，示出了本发明的一种实时驱动虚拟人的装置实施例的结构框图，具体可以包括：

[0124] 数据获取模块501，用于获取用于驱动虚拟人的待处理数据，所述待处理数据包括文本数据和语音数据中的至少一种；

[0125] 数据处理模块502，用于使用端到端模型对所述待处理数据进行处理，确定出所述待处理数据对应的手势特征序列；

[0126] 虚拟人驱动模块503，用于将所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人；

[0127] 其中，所述数据处理模块，用于获取所述待处理数据的文本特征和时长特征；根根据所述文本特征和所述时长特征，确定出所述手势特征序列。

[0128] 在一种可选实施方式中，数据处理模块502，用于通过fastspeech模型获取所述文本特征；通过时长模型获取所述时长特征，其中，所述时长模型为深度学习模型。

[0129] 在一种可选实施方式中，数据处理模块502，所述fastspeech模型输出面部特征序列和手势特征序列，用于所述文本特征和所述时长特征输入到所述fastspeech模型中，得到所述面部特征序列和所述手势特征序列。

[0130] 在一种可选实施方式中，虚拟人驱动模块503，用于将将所述面部特征序列和所述手势特征序列进行融合，得到融合特征序列；将所述融合特征序列输入到所述肌肉模型中。

[0131] 在一种可选实施方式中，虚拟人驱动模块503，用于基于所述时长特征，将所述面部特征序列和所述手势特征序列进行融合，得到所述融合特征序列。

[0132] 在一种可选实施方式中，所述面部特征序列对应的面部特征包括表情特征和唇部特征。

[0133] 对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0134] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0135] 关于上述实施例中的装置，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

[0136] 图6是根据一示例性实施例示出的一种用于实时驱动虚拟人的装置作为设备时的结构框图。例如，装置900可以是移动来电，计算机，数字广播终端，消息收发设备，游戏控制台，平板设备，医疗设备，健身设备，个人数字助理等。

[0137] 参照图6，装置900可以包括以下一个或多个组件：处理组件902，存储器904，电源组件906，多媒体组件908，音频组件910，输入/输出(I/O)的接口912，传感器组件914，以及通信组件916。

[0138] 处理组件902通常控制装置900的整体操作，诸如与显示，来电呼叫，数据通信，相机操作和记录操作相关联的操作。处理元件902可以包括一个或多个处理器920来执行指令，以完成上述的方法的全部或部分步骤。此外，处理组件902可以包括一个或多个模块，便于处理组件902和其他组件之间的交互。例如，处理组件902可以包括多媒体模块，以方便多媒体组件908和处理组件902之间的交互。

[0139] 存储器904被配置为存储各种类型的数据以支持在设备900的操作。这些数据的示例包括用于在装置900上操作的任何应用程序或方法的指令，联系人数据，来电簿数据，消息，图片，视频等。存储器904可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器(SRAM)，电可擦除可编程只读存储器(EEPROM)，可擦除可编程只读存储器(EPROM)，可编程只读存储器(PROM)，只读存储器(ROM)，磁存储器，快闪存储器，磁盘或光盘。

[0140] 电源组件906为装置900的各种组件提供电力。电源组件906可以包括电源管理系统，一个或多个电源，及其他与为装置900生成、管理和分配电力相关联的组件。

[0141] 多媒体组件908包括在所述装置900和用户之间的提供一个输出接口的屏幕。在一些实施例中，屏幕可以包括液晶显示器(LCD)和触摸面板(TP)。如果屏幕包括触摸面板，屏幕可以被实现为触摸屏，以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动运动动作的边界，而且还检测与所述触摸或滑动操作相关的持续时间和压力。在一些实施例中，多媒体组件908包括一个前置摄像头和/或后置摄像头。当设备900处于操作模式，如拍摄模式或视频模式时，前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。

[0142] 音频组件910被配置为输出和/或输入音频信号。例如，音频组件910包括一个麦克风(MIC)，当装置900处于操作模式，如呼叫模式、记录模式和语音识别模式时，麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器904或经由通信组件916发送。在一些实施例中，音频组件910还包括一个扬声器，用于输出音频信号。

[0143] I/O接口912为处理组件902和外围接口模块之间提供接口，上述外围接口模块可以是键盘，点击轮，按钮等。这些按钮可包括但不限于：主页按钮、音量按钮、启动按钮和锁定按钮。

[0144] 传感器组件914包括一个或多个传感器，用于为装置900提供各个方面的状态评估。例如，传感器组件914可以检测到设备900的打开/关闭状态，组件的相对定位，例如所述组件为装置900的显示器和小键盘，传感器组件914还可以检测装置900或装置900一个组件的位置改变，用户与装置900接触的存在或不存在，装置900方位或加速/减速和装置900的温度变化。传感器组件914可以包括接近传感器，被配置用来在没有任何的物理接触时检测

附近物体的存在。传感器组件914还可以包括光传感器，如CMOS或CCD图像传感器，用于在成像应用中使用。在一些实施例中，该传感器组件914还可以包括加速度传感器，陀螺仪传感器，磁传感器，压力传感器或温度传感器。

[0145] 通信组件916被配置为便于装置900和其他设备之间有线或无线方式的通信。装置900可以接入基于通信标准的无线网络，如WiFi，2G或3G，或它们的组合。在一个示例性实施例中，通信部件916经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中，所述通信部件916还包括近场通信(NFC)模块，以促进短程通信。例如，在NFC模块可基于射频识别(RFID)技术，红外数据协会(IrDA)技术，超宽带(UWB)技术，蓝牙(BT)技术和其他技术来实现。

[0146] 在示例性实施例中，装置900可以被一个或多个应用专用集成电路(ASIC)、数字信号处理器(DSP)、数字信号处理设备(DSPD)、可编程逻辑器件(PLD)、现场可编程门阵列(FPGA)、控制器、微控制器、微处理器或其他电子元件实现，用于执行上述方法。

[0147] 在示例性实施例中，还提供了一种包括指令的非临时性计算机可读存储介质，例如包括指令的存储器904，上述指令可由装置900的处理器920执行以完成上述方法。例如，所述非临时性计算机可读存储介质可以是ROM、随机存取存储器(RAM)、CD-ROM、磁带、软盘和光数据存储设备等。

[0148] 图7是本发明的一些实施例中服务器的结构框图。该服务器1900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，CPU)1922(例如，一个或一个以上处理器)和存储器1932，一个或一个以上存储应用程序1942或数据1944的存储介质1930(例如一个或一个以上海量存储设备)。其中，存储器1932和存储介质1930可以是短暂存储或持久存储。存储在存储介质1930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1922可以设置为与存储介质1930通信，在服务器1900上执行存储介质1930中的一系列指令操作。

[0149] 服务器1900还可以包括一个或一个以上电源1926，一个或一个以上有线或无线网络接口1950，一个或一个以上输入输出接口1958，一个或一个以上键盘1956，和/或，一个或一个以上操作系统1941，例如Windows ServerTM，Mac OS XTM，UnixTM，LinuxTM，FreeBSDTM等等。

[0150] 一种非临时性计算机可读存储介质，当所述存储介质中的指令由装置(设备或者服务器)的处理器执行时，使得装置能够执行一种实时驱动虚拟人的方法，所述方法包括：确定待处理文本对应的时长特征；所述待处理文本涉及至少两种语言；依据所述时长特征，确定所述待处理文本对应的目标语音序列；依据所述时长特征，确定所述待处理文本对应的目标图像序列；所述目标图像序列为依据文本样本及其对应的图像样本得到；所述文本样本对应的语言包括：所述待处理文本涉及的所有语言；对所述目标语音序列和所述目标图像序列进行融合，以得到对应的目标视频。

[0151] 本说明书是参照根据本说明书实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器

以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的设备。

[0152] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令设备的制造品，该指令设备实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0153] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0154] 尽管已描述了本说明书的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本说明书范围的所有变更和修改。

[0155] 显然，本领域的技术人员可以对本说明书进行各种改动和变型而不脱离本说明书的精神和范围。这样，倘若本说明书的这些修改和变型属于本说明书权利要求及其等同技术的范围之内，则本说明书也意图包含这些改动和变型在内。

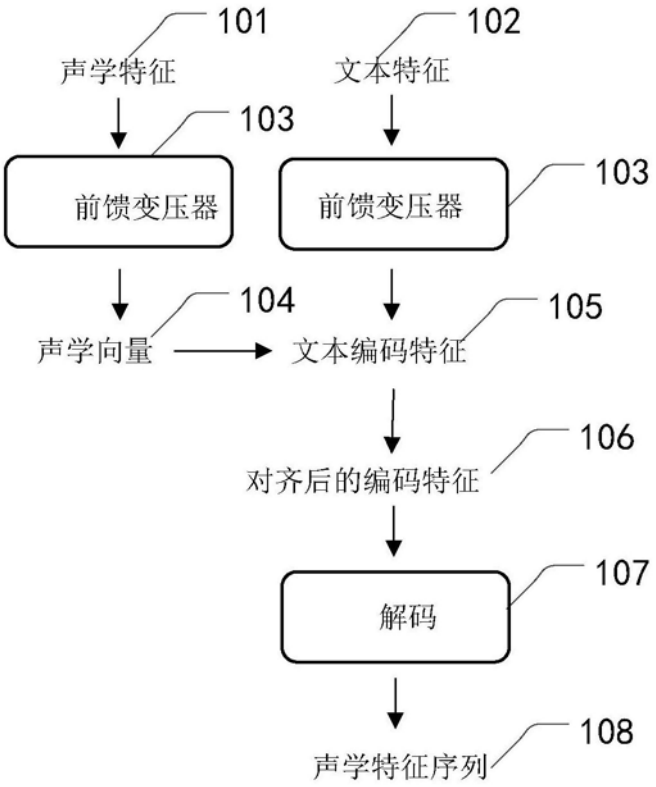

101
声学特征
102
文本特征
103
前馈变压器
103
前馈变压器
104
声学向量
105
文本编码特征
106
对齐后的编码特征
107
解码
108
声学特征序列

图1

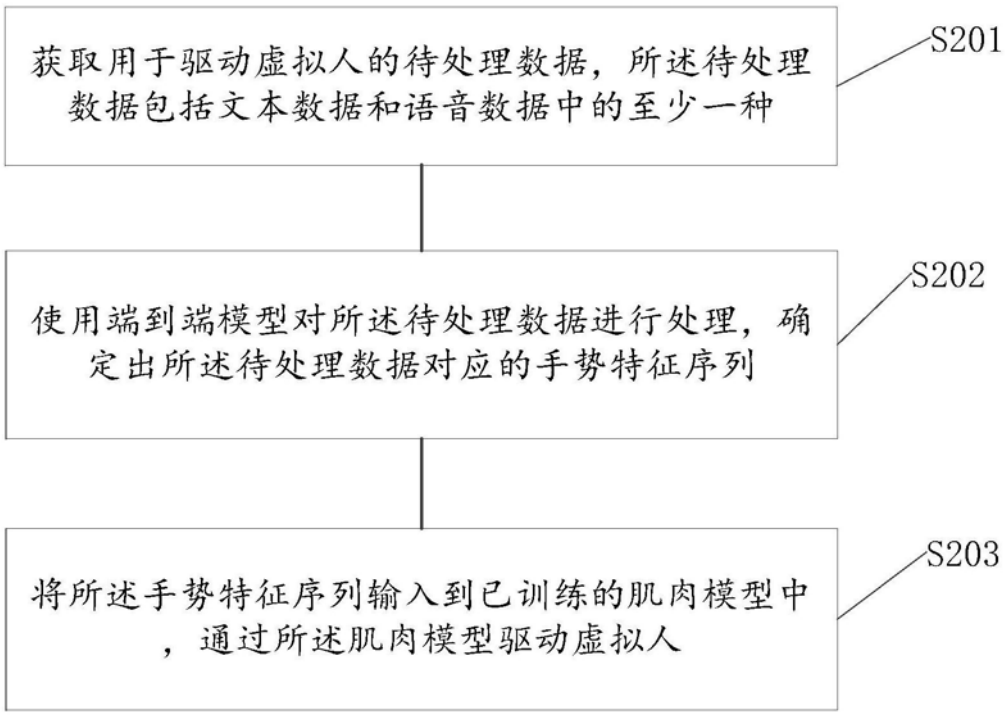

获取用于驱动虚拟人的待处理数据，所述待处理数据包括文本数据和语音数据中的至少一种
S201
使用端到端模型对所述待处理数据进行处理，确定出所述待处理数据对应的手势特征序列
S202
将所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人
S203

图2

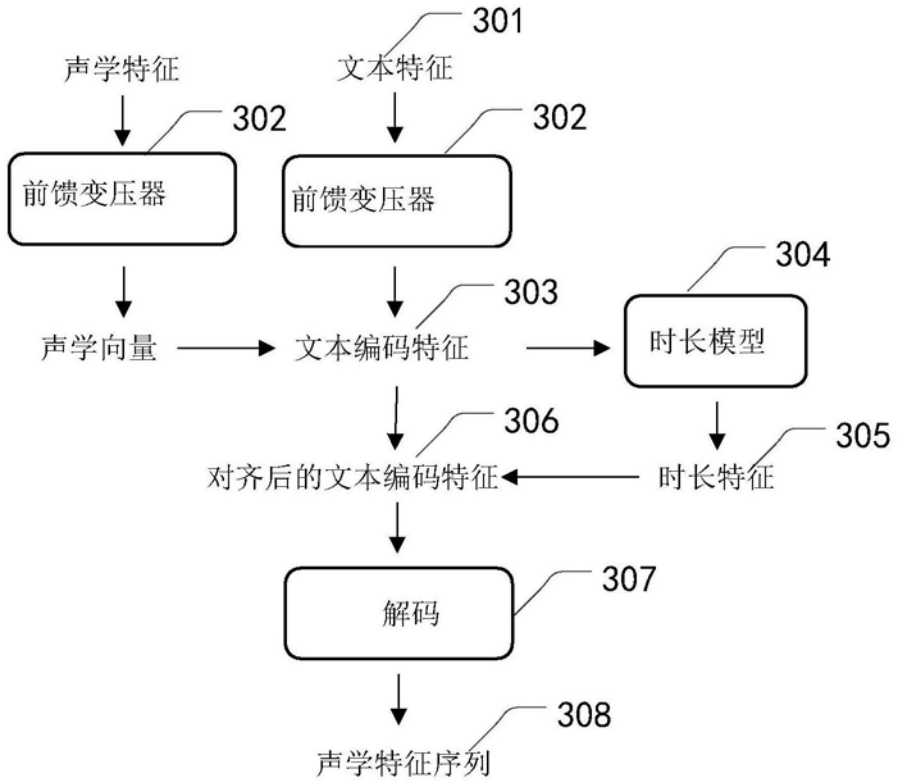

301
声学特征
文本特征
302
302
前馈变压器
前馈变压器
304
303
声学向量
文本编码特征
时长模型
306
305
对齐后的文本编码特征
时长特征
307
解码
308
声学特征序列

图3

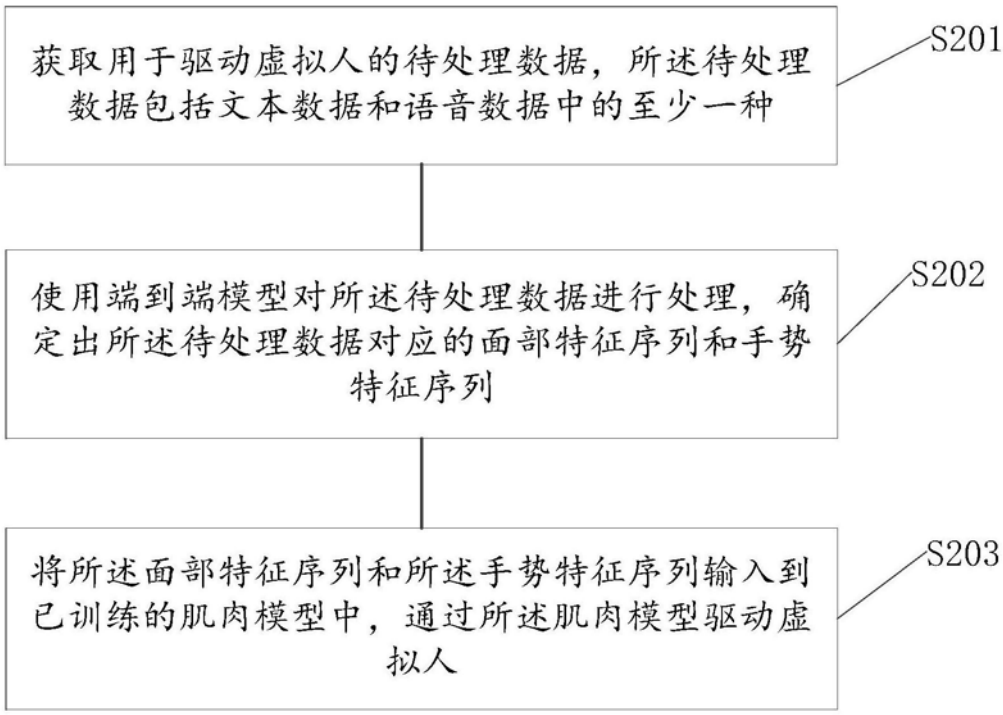

获取用于驱动虚拟人的待处理数据，所述待处理数据包括文本数据和语音数据中的至少一种
S201
使用端到端模型对所述待处理数据进行处理，确定出所述待处理数据对应的面部特征序列和手势特征序列
S202
将所述面部特征序列和所述手势特征序列输入到已训练的肌肉模型中，通过所述肌肉模型驱动虚拟人
S203

图4

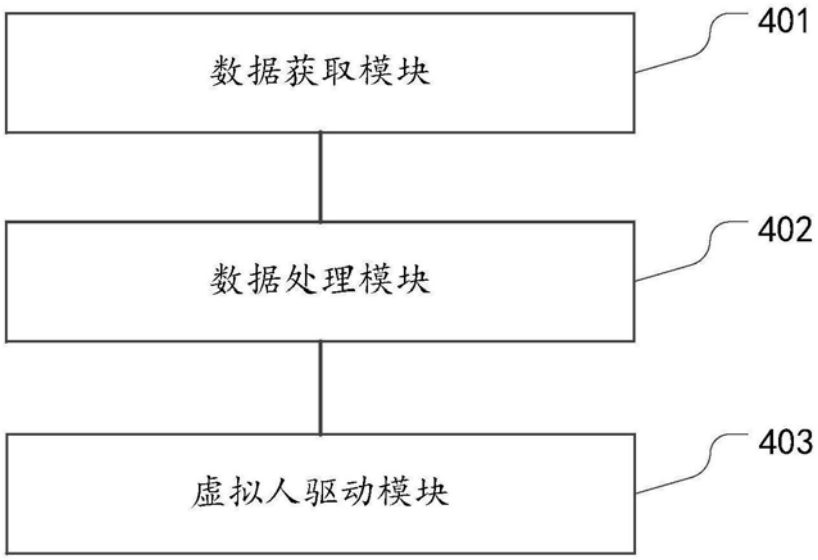


图5

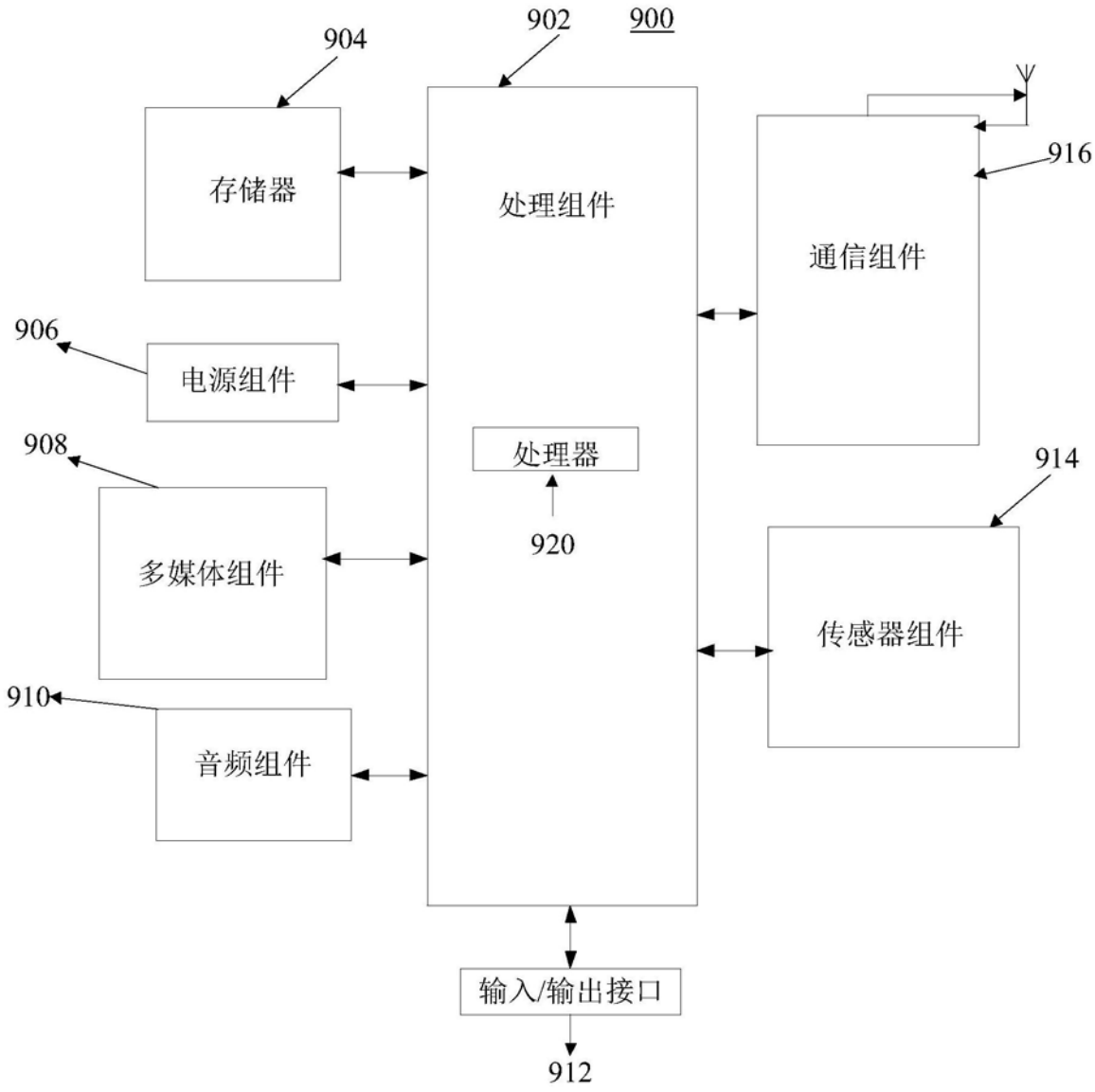

902
900
904
存储器
处理组件
通信组件
916
906
电源组件
处理器
908
920
914
多媒体组件
传感器组件
910
音频组件
输入/输出接口
912

图6

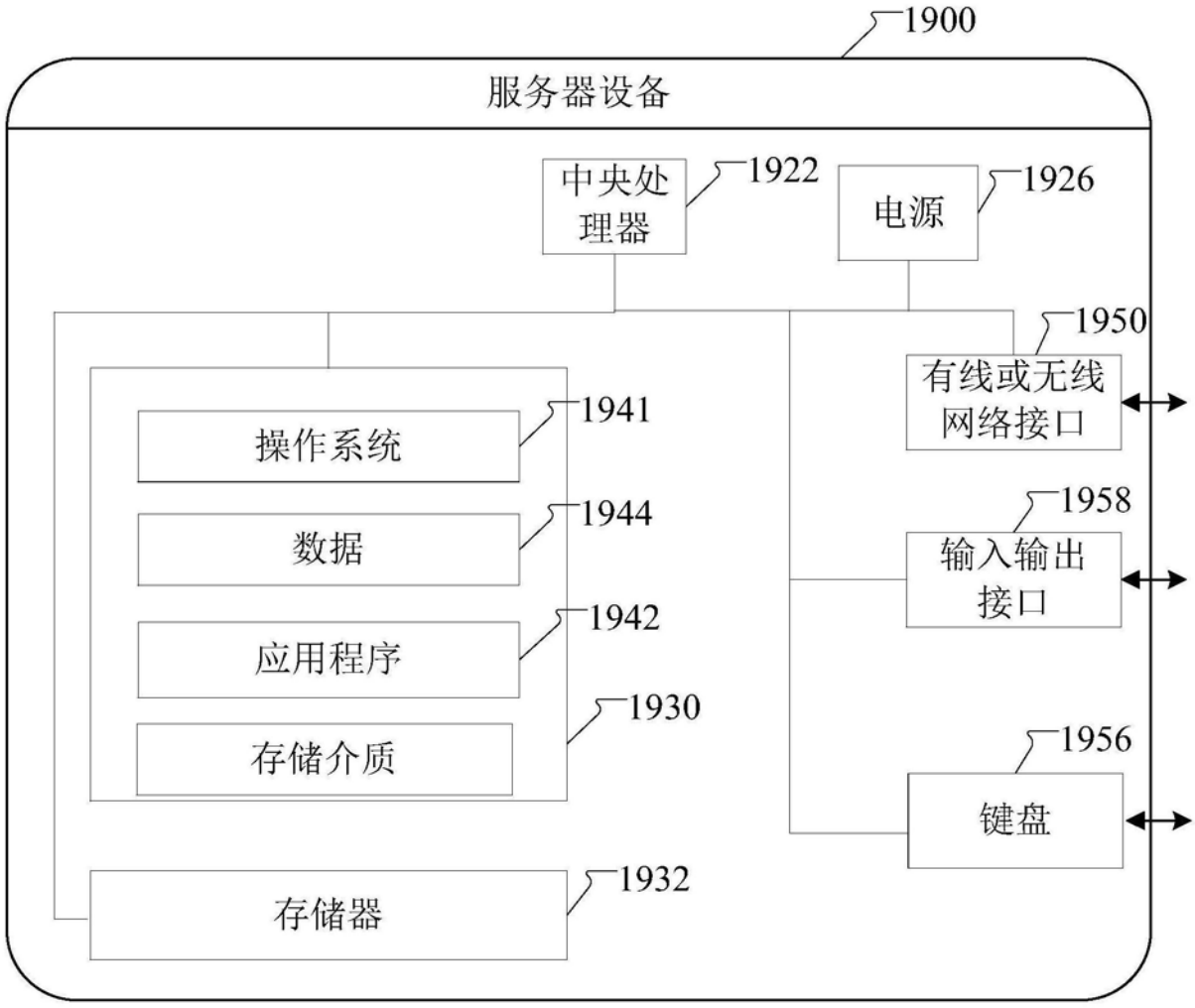

1900
服务器设备
中央处理器
1922
电源
1926
1950
有线或无线网络接口
操作系统
1941
数据
1944
1958
输入输出接口
应用程序
1942
存储介质
1930
1956
键盘
存储器
1932

图7