

(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
2023 年 5 月 25 日 (25.05.2023)



(10) 国际公布号
WO 2023/088080 A1

(51) 国际专利分类号:
G10L 15/26 (2006.01) **H04N 21/234** (2011.01)
G06N 3/04 (2023.01)

(21) 国际申请号: PCT/CN2022/128584

(22) 国际申请日: 2022 年 10 月 31 日 (31.10.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111386695.8 2021 年 11 月 22 日 (22.11.2021) CN

(71) 申请人: 上海商汤智能科技有限公司 (SHANGHAI SENSETIME INTELLIGENT TECHNOLOGY CO., LTD.) [CN/CN]; 中国上海

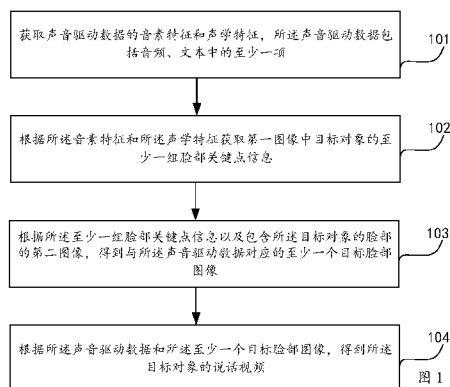
市徐汇区桂平路 391 号 3 号楼 1605A 室, Shanghai 200233 (CN)。

(72) 发明人: 王宇欣 (WANG, Yuxin); 中国上海市徐汇区桂平路 391 号 3 号楼 1605A 室, Shanghai 200233 (CN)。 吴文岩 (WU, Wenyan); 中国上海市徐汇区桂平路 391 号 3 号楼 1605A 室, Shanghai 200233 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区上地三街 9 号嘉华大厦 B 座 409, Beijing 100085 (CN)。

(54) **Title:** SPEAKING VIDEO GENERATION METHOD AND APPARATUS, AND ELECTRONIC DEVICE AND STORAGE MEDIUM

(54) 发明名称: 说话视频生成方法、装置、电子设备以及存储介质



- 101 Acquire phoneme features and acoustic features of voice drive data, wherein the voice drive data comprises at least one of audio and text
102 Acquire at least one set of facial key point information of a target object in a first image according to the phoneme features and the acoustic features
103 According to the at least one set of facial key point information and a second image including the face of the target object, obtain at least one target facial image corresponding to the voice drive data
104 Obtain a speaking video of the target object according to the voice drive data and the at least one target facial image

(57) **Abstract:** Disclosed are a speaking video generation method and apparatus, and a device and a storage medium. The method comprises: acquiring phoneme features and acoustic features of voice drive data, wherein the voice drive data comprises at least one of audio and text; acquiring at least one set of facial key point information of a target object in a first image according to the phoneme features and the acoustic features; according to the at least one set of facial key point information and a second image including the face of the target object, obtaining at least one target facial image corresponding to the voice drive data, wherein a set area comprising a specific part of the target object in the second image is blocked; and obtaining a speaking video of the target object according to the voice drive data and the at least one target facial image.

(57) **摘要:** 公开了一种说话视频生成方法、装置、设备以及存储介质, 所述方法包括: 获取声音驱动数据的音素特征以及声学特征, 所述声音驱动数据包括音频、文本中的至少一项; 根据所述音素特征和所述声学特征获取第一图像中目标对象的至少一组脸部关键点信息; 根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像, 得到与所述声音驱动数据对应的至少一个目标脸部图像, 其中, 所述第二图像中包括所述目标对象的特定部位的设定区域被遮挡; 根据所述声音驱动数据和所述至少一个目标脸部图像, 得到所述目标对象的说话视频。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

- 关于申请人有权要求在先申请的优先权 (细则4.17(iii))

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

说话视频生成方法、装置、电子设备以及存储介质

相关申请的交叉引用

本专利申请要求于 2021 年 11 月 22 日提交的、申请号为 202111386695.8 的中国专利申请的优先权，其全部内容通过引用并入本文中。

技术领域

[01]本公开涉及计算机视觉技术领域，具体涉及一种说话视频生成方法、装置、设备以及存储介质。

背景技术

[02]说话视频生成技术是语音驱动人物形象、跨模态视频生成等中用到的一类重要技术，在虚拟数字对象商业化中起到关键作用。目前通常根据语音帧确定对应的口型图像，从而获取输出语音对应的一系列口型图像来生成说话视频，然而该方法所生成的视频中说话人的口型准确度较低且口型变化生硬。

发明内容

[03]本公开实施例提供一种说话视频生成方案。

[04]根据本公开的第一方面，提供一种说话视频生成方法，所述方法包括：获取声音驱动数据的音素特征以及声学特征，所述声音驱动数据包括音频、文本中的至少一项；根据所述音素特征和所述声学特征获取第一图像中目标对象的至少一组脸部关键点信息；根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像，其中，所述第二图像中包括所述目标对象的特定部位的设定区域被遮挡；根据所述声音驱动数据和所述至少一个目标脸部图像，得到所述目标对象的说话视频。

[05]结合本公开提供的任一实施方式，所述获取声音驱动数据的音素特征以及声学特征，包括：获取所述声音驱动数据对应的音频所包含的音素以及各个音素对应的时间戳，得到所述声音驱动数据的音素特征；对所述声音驱动数据对应的音频进行特征提取，得到所述声音驱动数据的声学特征。

[06]结合本公开提供的任一实施方式，所述根据所述音素特征和所述声学特征获取第一图像中目标对象的至少一组脸部关键点信息，包括：获取所述音素特征所包含的多个子音素特征以及所述多个子音素特征对应的子声学特征；将所述子音素特征和对应的子声学特征输入至脸部关键点提取网络，得到与所述子音素特征和所述子声学特征对应的脸部关键点信息。

[07]结合本公开提供的任一实施方式，所述脸部关键点信息包括 3D 脸部关键点信息，在根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像之前，所述方法还包括：将所述 3D 脸部关键点信息投影到 2D 平面上，得到所述 3D 脸部关键点信息对应的 2D 脸部关键点信息；利用所述 2D 脸部关键点信息更新所述脸部关键点信息。

[08]结合本公开提供的任一实施方式，在根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像之前，所述方法还包括：对多组脸部关键点信息进行滤波处理，使每个图像帧的脸部关

键点信息与该图像帧的相邻帧的脸部关键点信息之间的变化量满足设定条件。

[09]结合本公开提供的任一实施方式，所述根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像，包括：将每组脸部关键点信息与所述第二图像输入至脸部补全网络，得到与所述脸部关键点信息对应的目标脸部图像，其中，所述脸部补全网络用于根据脸部关键点信息对所述第二图像中被遮挡的设定区域进行补全。

[10]结合本公开提供的任一实施方式，所述根据所述声音驱动数据和所述至少一个目标脸部图像，得到所述目标对象的说话视频，包括：将所述至少一个目标脸部图像与设定背景图像进行融合，得到第一图像序列；根据所述第一图像序列与所述声音驱动数据对应的音频，得到所述目标对象的说话视频。

[11]结合本公开提供的任一实施方式，所述脸部关键点提取网络利用音素特征样本和对应的声学特征样本训练得到，其中，所述音素特征样本和所述声学特征样本包括标注的脸部关键点信息。

[12]结合本公开提供的任一实施方式，所述脸部关键点提取网络通过以下方式训练得到：根据所述音素特征样本和对应的声学特征样本，对初始脸部关键点提取网络进行训练，在网络损失的变化满足收敛条件时完成训练得到所述脸部关键点提取网络，其中，所述网络损失包括所述初始神经网络预测得到的脸部关键点信息与标注的脸部关键点信息之间的差异。

[13]结合本公开提供的任一实施方式，所述音素特征样本和所述声学特征样本通过对一对对象的音频的音素特征和声学特征进行所述对象的脸部关键点信息标注得到。

[14]结合本公开提供的任一实施方式，所述音素特征样本和所述声学特征样本通过以下方式得到：获取所述对象的说话视频；根据所述说话视频获取多个脸部图像，以及与每个所述脸部图像对应的至少一个音频帧；获取每个所述脸部图像对应的至少一个音频帧的音素特征以及声学特征；根据所述多个脸部图像获取脸部关键点信息，并根据所述脸部关键点信息对所述音素特征和所述声学特征进行标注，得到所述音素特征样本和所述声学特征样本。

[15]结合本公开提供的任一实施方式，所述脸部补全网络利用生成对抗网络训练得到，所述生成对抗网络包括所述脸部补全网络和第一鉴别网络，所述训练的网络损失包括：第一损失，用于指示所述脸部补全网络输出的脸部补全图像与完整脸部图像之间的差异，其中，所述完整脸部图像是所述脸部关键点信息对应的脸部图像；第二损失，用于指示所述第一鉴别网络对于输入图像输出的分类结果与所述输入图像的标注信息之间的差异，其中，所述标注信息指示所述输入图像为所述脸部补全网络输出的脸部补全图像或者为真实脸部图像。

[16]结合本公开提供的任一实施方式，所述生成对抗网络还包括第二鉴别网络，所述训练的网络损失还包括：第三损失，用于指示所述第二鉴别网络对于所述脸部补全图像与音素特征之间的对应的判别结果与真实对应结果之间的差异。

[17]根据本公开的第二方面，提供一种说话视频生成装置，所述装置包括：第一获取单元，用于获取声音驱动数据的音素特征以及声学特征，所述声音驱动数据包括音频、文本中的至少一项；第二获取单元，用于根据所述音素特征和所述声学特征获取第一图像中目标对象的至少一组脸部关键点信息；第一得到单元，用于根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像，其中，所述第二图像中包括所述目标对象的特定部位的设定区域被遮挡；第二得到单元，用于根据所述声音驱动数据和所述至少一个目标脸部图像，得

到所述目标对象的说话视频。

[18]根据本公开的第三方面，提供一种电子设备，所述设备包括存储器、处理器，所述存储器用于存储可在处理器上运行的计算机指令，所述处理器用于在执行所述计算机指令时实现本公开提供的任一实施方式所述的说话视频生成方法。

[19]根据本公开的第四方面，提供一种计算机可读存储介质，其上存储有计算机程序，所述程序被处理器执行时实现本公开提供的任一实施方式所述的说话视频生成方法。

[20]本公开一个或多个实施例的说话视频生成方法、装置、设备及计算机可读存储介质，根据声音驱动数据的音素特征以及声学特征，获取第一图像中目标对象的至少一组脸部关键点信息；并根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像，其中，所述第二图像中包括所述目标对象的特定部位的设定区域被遮挡；最后根据所述声音驱动数据和所述至少一个目标脸部图像，得到所述目标对象的说话视频。本公开实施例根据与声音驱动数据对应的目标对象的脸部关键信息以及所述目标对象的遮挡了特定部位的图像来生成目标脸部图像，所得到的所述目标对象的说话视频中所述目标对象的口型与所述声音驱动数据的匹配度高，且口型变化连贯，所述目标对象说话状态真实、自然。

附图说明

[21]为了更清楚地说明本说明书一个或多个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说明书一个或多个实施例中记载的一些实施例，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。

[22]图1是本公开至少一个实施例提出的一种说话视频生成方法的流程图；

[23]图2是本公开至少一个实施例提出的脸部关键点提取网络训练方法流程图；

[24]图3是本公开至少一个实施例提出的样本获取方法流程图；

[25]图4是本公开至少一个实施例提出的另一种说话视频生成方法的流程图；

[26]图5是图4所示的说话视频生成方法的示意图；

[27]图6是本公开至少一个实施例提出的说话视频生成方法中获取脸部关键点信息的示意图；

[28]图7是本公开至少一个实施例提出的说话视频生成装置的结构示意图；

[29]图8是本公开至少一个实施例提出的电子设备的结构示意图。

具体实施方式

[30]这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本公开相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本公开的一些方面相一致的装置和方法的例子。

[31]本文中术语“和/或”，仅仅是一种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：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本文中术语“至少一种”表示多种中的任意一种或多种中的至少两种的任意组合，例如，包括A、B、C中的至少一种，可以表示包括从A、B和C构成的集合中选择的

任意一个或多个元素。

[32]本公开至少一个实施例提供了一种说话视频生成方法，该方法可以由终端设备或服务器等电子设备执行。所述终端设备可以是固定终端或移动终端，例如手机、平板电脑、游戏机、台式机、广告机、一体机、车载设备等等，所述服务器包括本地服务器或云端服务器等。所述方法还可以通过处理器调用存储器中存储的计算机可读指令的方式来实现。

[33]图 1 示出根据本公开至少一个实施例的说话视频生成方法的流程图，如图 1 所示，所述方法包括步骤 101~步骤 104。

[34]在步骤 101 中，获取声音驱动数据的音素特征以及声学特征。

[35]音素是构成音节的最小语音单位，所述声音驱动数据对应的音频可以包含一个或多个音素，音素特征可以包括表示各个音素的发音起止时间的特征。以所述音频是“你好”语音段为例，所述声音驱动数据的音素特征例如可以包括： $n[0, 0.2]$ ， $i3[0.2, 0.4]$ ， $h[0.5, 0.7]$ ， $ao3[0.7, 1.2]$ ，其中，[]内的数据指示相应音素的发音起止时间，单位例如为秒。在本公开实施例中，可以通过获取所述声音驱动数据对应的音频所包含的音素以及各个音素对应的时间戳，得到所述声音驱动数据的音素特征。

[36]声学特征主要用于描述音频的发音特点，所述声学特征包括但不限于线性预测参数、梅尔频率倒谱系数、感知线性高效能系数等中的至少一种。在本公开实施例中，所述声学特征例如为梅尔频率倒谱系数。在本公开实施例中，所述声音驱动数据的声学特征可以通过对所述声音驱动数据对应的音频进行特征提取得到。

[37]所述声音驱动数据被预先存储在执行该说话视频生成方法的电子设备中，或预先存储在该电子设备以外的其他设备中，或者通过声音采集设备现场采集获得，等等，本公开不对所述声音驱动数据的来源进行限制。在本公开实施例，所述声音驱动数据可以包括音频、文本中的至少一项。

[38]在所述声音驱动数据只包括音频的情况下，可以通过对所述音频进行语音识别，确定所述音频对应的文本（例如，文字信息）；

[39]在所述声音驱动数据只包括文本的情况下，可以通过对所述文本进行语音合成，将文本所对应的文字信息转化为音频（例如，语音段）；

[40]在所述声音驱动数据包括音频和文本的情况下，该音频和文本对应于相同的发音。例如，该文本是“你好”，所述声音驱动数据中的音频则是发出“你好”声音的语音段。

[41]在一些实施例中，通过对所述声音驱动数据对应的音频和文本进行对齐操作，可以得到所述声音驱动数据的音素特征。其中，对齐操作是指将音频中各个语音段与该语音段所对应的文本中的音素进行对齐，也即确定在所述音频中从何时开始发出文本对应的读音。通过对音频和文本进行对齐操作，一方面确定了所述音频所包含的音素，同时根据读音的持续时间可以得到各个音素对应的时间戳，从而可以得到所述声音驱动数据的音素特征。

[42]仍以“你好”为例，在对音频和文本进行对齐操作后，则可以确定在 0~0.2 秒发出音素“n”的读音，在 0.2~0.4 秒发出“i3”的读音，等等，从而可以得到所述声音驱动数据的音素特征。本领域技术人员应当理解，也可以通过其他方式获取所述声音驱动数据的音素特征，本公开实施例对此不进行限制。

[43]在本公开实施例中，通过利用音素特征，可以采用音频或文本任意一种方式来对目标对象进行驱动以生成目标对象的说话视频。

[44]在步骤 102 中, 根据所述音素特征和所述声学特征获取第一图像中目标对象的至少一组脸部关键点信息。

[45]目标对象(例如, 人)在发出不同的语音时, 口型会发生相应的变化, 因此, 目标对象的嘴部区域或者包含嘴部区域的设定区域内的脸部关键点的位置会发生相应的变化。由此可知, 一个音频帧的音素特征和声学特征是与目标对象的一组脸部关键点信息相对应的。目标对象在发出某一音素的读音时, 其面部对应的脸部关键点信息是可以确定的。其中, 所述脸部关键点信息包括目标对象的脸部关键点(例如, 五官和脸部轮廓对应的关键点)在包含所述目标对象的脸部的图像(例如, 第一图像)中的位置信息。在本公开中, 可以将各个脸部关键点在同一时刻的信息称为一组脸部关键点信息。

[46]以生成第一图像中目标对象的说话视频为例, 在本步骤中, 根据所述声音驱动数据的音素特征和声学特征, 获取第一图像中目标对象的至少一组脸部关键点信息。在所述声音驱动数据对应的音频包括多个音素的情况下, 可以得到这些音素和相应的声学特征所对应的脸部关键点信息序列。其中, 所述脸部关键点信息序列包括按照时间顺序排列的多组脸部关键点信息。

[47]本公开实施例在所述声音驱动数据的音素特征的基础上还利用了声学特征, 使得所获取的脸部关键点信息与该声音驱动数据对应的音频的发音特征更加匹配, 使得后续生成的说话视频更加真实。

[48]在步骤 103 中, 根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像, 得到与所述声音驱动数据对应的至少一个目标脸部图像。

[49]其中, 所述第二图像是包含所述目标对象的脸部的图像, 所述第二图像可以通过对所述第一图像进行遮挡处理获得, 也可以通过对与所述第一图像不同的另一图像进行遮挡处理获得。例如, 第一图像是目标对象 A 正在微笑的脸部图像, 与第一图像不同的另一图像可以是目标对象 A 正在撇嘴的脸部图像。

[50]所述第二图像中的包括所述目标对象的特定部位(例如, 嘴部)的设定区域被遮挡, 所述设定区域包括目标对象在说话时脸部关键点的位置会发生变化的区域, 例如, 可以是所述目标对象的脸部的下半部分, 也可以是额头以下的脸部区域, 还可以是嘴部区域, 本公开实施例对于遮挡的具体区域不进行限制。

[51]在一些实施例中, 可以通过对所述设定区域进行噪声填充, 生成所述设定区域被遮挡的第二图像。其中, 对所述设定区域进行噪声填充是指利用随机生成的像素值对所述设定区域内的各个像素进行设置。本领域技术人员应当理解, 也可以通过其他方式进行所述设定区域的遮挡, 本公开对此不进行限制。

[52]根据在步骤 102 中所得到的至少一组脸部关键点信息, 可以对所述第二图像中被遮挡的部分(即, 设定区域)进行补全, 从而使得第二图像中被遮挡的设定区域中的脸部关键点的分布, 与声音驱动数据的音素特征和声学特征一致。这样, 根据所述至少一组脸部关键点信息以及所述第二图像生成的至少一个目标脸部图像中, 与所述设定区域相对应区域(即, 补全后的区域)的脸部关键点信息与所述声音驱动数据是匹配的。

[53]在步骤 104 中, 根据所述声音驱动数据和所述至少一个目标脸部图像, 得到所述目标对象的说话视频。

[54]本公开实施例中, 得到的所述目标对象的说话视频中, 所述目标对象输出的声音为所述声音驱动数据对应的音频, 而所述说话视频的各个图像帧中, 所述目标对象的脸部关键点信息是与输出的声音的音素特征和声学特征对应的。由此生成的说话视频中所述目标对象的口型和说话表情是与发音一致的, 使观众产生所述目标对象正在说话的感觉。

[55]本公开实施例根据与声音驱动数据对应的目标对象的至少一组脸部关键点信息以及所述目标对象的遮挡了特定部位的图像来生成至少一个目标脸部图像，所得到的所述目标对象的说话视频中所述目标对象的口型与所述声音驱动数据的匹配度高，且口型变化连贯，所述目标对象说话状态真实、自然。

[56]在一些实施例中，可以将所述至少一个目标脸部图像与设定背景图像进行融合，得到第一视频，并根据所述第一视频与所述声音驱动数据对应的音频，得到所述目标对象的说话视频。在一个示例中，可以将所述目标脸部图像中脸部区域的像素作为前景像素，与所述设定背景图像进行叠加，以实现所述目标脸部图像与设定背景图像的融合。本领域技术人员应当理解，可以采用多种方式对目标脸部图像与设定背景图像进行融合，本公开对此不进行限制。

[57]通过上述方法，可以生成目标对象在任意背景下的说话视频，丰富了说话视频生成方法的应用场景。

[58]在一些实施例中，可以利用脸部关键点提取网络，得到所述音素特征和所述声学特征对应的第一图像中目标对象的至少一组脸部关键点信息。

[59]首先，获取所述音素特征所包含的多个子音素特征以及所述多个子音素特征对应的子声学特征。

[60]在一个示例中，可以通过在所述声音驱动数据的音素特征和声学特征上进行滑窗的方式，获得所述音素特征所包含的多个子音素特征以及与所述多个子音素特征对应的子声学特征。比如，音素特征和声学特征可以按照时间窗口的长度来分别划分出多个子音素特征和子声学特征。具体地，在对所述声音驱动数据的音素特征和声学特征进行滑窗的过程中，可以将每次滑窗操作后获得的该次时间窗口内的音素特征和声学特征，作为子音素特征和子声学特征，并且，在相同时间窗口内的子音素特征和子声学特征是对应于同一语音段的。

[61]接下来，将所述子音素特征和对应的子声学特征输入至经训练的脸部关键点提取网络，得到与所述子音素特征和所述子声学特征对应的脸部关键点信息。具体的，可以将多个子音素特征和对应的多个子声学特征，以多个子音素特征-子声学特征对的形式，按时间顺序输入至脸部关键点提取网络。所述脸部关键点提取网络用于根据每个子音素特征-子声学特征对，确定对应的一组脸部关键点信息。在将所有的子音素特征-子声学特征对输入至脸部关键点提取网络后，即可以得到所述声音驱动数据对应的多组脸部关键点信息。

[62]在本公开实施例中，通过经训练的脸部关键点提取网络，得到与各个子音素特征-子声学特征对对应的脸部关键点信息，可以实现所述目标对象的发音与口型和说话表情的良好匹配。

[63]在本公开实施例中，所述脸部关键点提取网络可以是三维 3D 脸部关键点提取网络，也即所输出的脸部关键点信息是 3D 脸部关键点信息，所述 3D 脸部关键点信息除了包含所述脸部关键点的位置信息外，还包括所述脸部关键点的深度信息；所述脸部关键点提取网络也可以是二维 2D 脸部关键点提取网络，也即所输出的脸部关键点信息是 2D 脸部关键点信息。

[64]在所述脸部关键点信息为 3D 脸部关键点信息的情况下，在根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像之前，所述方法还包括：将所述 3D 脸部关键点信息投影到 2D 平面上，得到所述 3D 脸部关键点信息对应的 2D 脸部关键点信息；利用所述 2D 脸部关键点信息更新所述脸部关键点信息。之后，根据至少一组 2D 脸部关键点信息以及包含

所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像；最后，根据所述声音驱动数据和所述至少一个目标脸部图像，得到所述目标对象的说话视频。

[65]在一些实施例中，可以对多组脸部关键点信息进行滤波处理，使最后获得的说话视频中每个图像帧的脸部关键点信息与该图像帧的相邻帧（包括前一帧和/或后一帧）的脸部关键点信息之间的变化量满足设定条件，该设定条件例如可以包括一个图像帧中各个脸部关键点的位置与相邻帧中对应脸部关键点的位置之间的变化量均小于设定阈值。通过上述方法可以滤除脸部关键点信息变化幅度较大的抖动帧，避免所生成的说话视频中出现口型突然变化的情况。

[66]在一个示例中，可以通过对多组脸部关键点信息在时间窗口上进行高斯滤波，来实现对所述多组脸部关键点信息所对应的连续帧的移动平均处理。其中，移动平均处理指对每一帧的脸部关键点的值和相邻帧的脸部关键点的值进行加权平均，利用加权平均的结果更新该帧的脸部关键点的值。

[67]在一些实施例中，可以通过以下方式得到与所述声音驱动数据对应的至少一个目标脸部图像：将每组脸部关键点信息与包含所述目标对象的脸部的第二图像输入至脸部补全网络，得到与所述脸部关键点信息对应的目标脸部图像，其中，所述脸部补全网络用于根据脸部关键点信息对所述第二图像中被遮挡的设定区域进行补全。

[68]在本公开实施例中，通过脸部补全网络，根据脸部关键点信息对所述第二图像中被遮挡的设定区域进行补全，可以使得所述设定区域的脸部关键点信息与输入的脸部关键点信息一致，从而使得所述目标对象的口型以及说话表情与发出的语音匹配，并且利用脸部补全网络对所述第二图像中被遮挡的设定区域进行补全，可以生成清晰程度高的目标脸部图像。

[69]在一些实施例中，所述脸部关键点提取网络可以利用音素特征样本和声学特征样本进行训练得到。该训练方法可以由服务器执行，并且执行该训练方法的服务器与执行上述说话视频生成方法的设备可以是不同的。

[70]图 2 示出了本公开至少一个实施例提出的脸部关键点提取网络的训练方法，如图 2 所示，该训练方法包括步骤 201~202。

[71]在步骤 201 中，获取音素特征样本和对应的声学特征样本，所述音素特征样本和所述声学特征样本包括标注的所述目标对象的脸部关键点信息。其中，所述音素特征样本和对应的声学特征样本，是基于同一语音段得到的，并且所述音素特征样本和对应的声学特征样本中所标注的脸部关键点信息是相同的。

[72]在步骤 202 中，根据所述音素特征样本和对应的声学特征样本，对初始脸部关键点提取网络进行训练，在网络损失的变化满足收敛条件时完成训练得到所述脸部关键点提取网络，其中，所述网络损失包括所述初始神经网络预测得到的脸部关键点信息与标注的脸部关键点信息之间的差异。

[73]在一些实施例中，所述音素特征样本和所述声学特征样本通过对一对象的音频的音素特征和声学特征进行所述对象的脸部关键点信息标注得到。一个示例中，可以通过图 3 所示的方法获取所述音素特征样本和对应的声学特征样本。

[74]在步骤 301 中，获取所述对象的说话视频。其中，该对象可以是待生成说话视频的上述目标对象，也可以是与该目标对象不同的对象。

[75]在一示例中，在想要生成某一目标对象的说话视频的情况下，则获取该目标对象的已有说话视频，用于获取音素特征样本和声学特征样本。

[76]在步骤 302 中, 根据所述说话视频获取多个脸部图像, 以及与每个所述脸部图像对应的至少一个音频帧。

[77]通过对所述说话视频进行拆分, 得到所述说话视频对应的语音段以及所述说话视频所包含的多个脸部图像。其中, 所述语音段中的多个音频帧与所述多个脸部图像具有对应关系。

[78]在步骤 303 中, 获取每个所述脸部图像对应的至少一个音频帧的音素特征以及声学特征。

[79]根据所述多个脸部图像与所述语音段中的多个音频帧的对应关系, 获取其中任意脸部图像对应的至少一个音频帧的音素特征以及声学特征。

[80]在步骤 304 中, 根据所述多个脸部图像获取脸部关键点信息, 并根据所述脸部关键点信息对所述音素特征和所述声学特征进行标注, 得到所述音素特征样本和所述声学特征样本。

[81]在本公开实施例中, 通过根据待生成说话视频所针对的目标对象的已有说话视频, 生成音素特征样本和声学特征样本, 可以准确地建立起所述目标对象所说的语音的音素特征和声学特征与脸部关键点信息之间的关联, 可以更好地实现对脸部关键点生成网络的训练。

[82]在一些实施例中, 所述脸部补全网络可以利用生成对抗网络进行训练。该训练方法可以由服务器执行, 并且执行该训练方法的服务器与执行上述说话视频生成方法的设备可以是不同的。

[83]其中, 所述生成对抗网络包括所述脸部补全网络和第一鉴别网络, 所述脸部补全网络用于根据脸部关键点信息对输入的遮挡脸部图像(即, 存在遮挡区域的脸部图像)进行补全, 生成脸部补全图像, 其中, 所述遮挡脸部图像是通过从完整脸部图像中包括特定部位(例如, 嘴部)的设定区域进行遮挡得到的, 所述完整脸部图像可以是所述脸部关键点信息对应的脸部图像; 所生成的脸部补全图像与真实脸部图像被随机输入至所述第一鉴别网络, 所述第一鉴别网络输出对于输入图像的鉴别结果, 也即判断所述输入图像是脸部补全图像或真实脸部图像。

[84]利用所述生成对抗网络对所述脸部补全网络进行训练的损失包括:

[85]第一损失, 用于指示所述脸部补全网络输出的脸部补全图像与完整脸部图像之间的差异, 其中, 所述完整脸部图像是所述脸部关键点信息对应的脸部图像;

[86]第二损失, 用于指示所述第一鉴别网络对于输入图像输出的分类结果与所述输入图像的标注信息之间的差异, 其中, 所述标注信息指示所述输入图像为所述脸部补全网络输出的脸部补全图像或者为真实脸部图像。

[87]在所述训练的损失的变化满足收敛条件时完成训练, 得到所述脸部补全网络。

[88]在本公开实施例中, 利用生成对抗网络对所述脸部补全网络进行训练, 可以提高所述脸部补全网络输出的脸部补全图像的准确度, 有利于提高所生成的所述目标对象的说话视频的图像质量。

[89]在一些实施例中, 还可以增加用于判断脸部补全图像是否与音素特征对齐的第二鉴别网络, 以辅助所述脸部补全网络的训练。在该训练方法中, 所述脸部补全网络输出的脸部补全图像输入至所述第二鉴别网络。

[90]该训练的损失在上述第一损失和第二损失之外, 还包括第三损失, 所述第三损失用于指示所述第二鉴别网络对于所述脸部补全图像与音素特征之间的对应(例如, 对齐)

的判别结果与真实对应结果之间的差异。

[91]通过增加第二鉴别网络对所述脸部补全网络进行训练，进一步提高了音素特征与脸部关键点的对齐效果，有利于提高说话视频的质量。

[92]以下结合图4所示的说话视频生成方法流程图、图5所示的说话视频生成方法示意图和图6所示的获取脸部关键点信息的示意图，对本公开实施例提出的一种说话视频生成方法进行描述。

[93]在步骤401中，对声音驱动数据对应的音频和文本进行对齐操作，得到所述声音驱动数据的音素特征。

[94]所述声音驱动数据对应的音频例如图6所示，为“你好”语音段，所述声音驱动数据对应的文本则为“你好”文本。通过对所述音频和文本进行对齐操作，得到所述声音驱动数据的音素特征。

[95]如图5所示，在所述声音驱动数据只包括音频的情况下，可以通过对所述音频进行语音识别，确定所述音频对应的文本；在所述声音驱动数据只包括文本的情况下，可以通过对所述文本进行语音合成，将文本所对应的文字信息转化为音频。

[96]在步骤402中，对所述声音驱动数据对应的音频进行特征提取，得到所述声音驱动数据的梅尔倒谱特征，也即梅尔频率倒谱系数。

[97]在步骤403中，通过在所述声音驱动数据的音素特征和声学特征上进行滑窗的方式，获得所述音素特征所包含的多个子音素特征以及与所述多个子音素特征对应的子声学特征。时间窗口如图6中的虚线框所示，箭头示出时间窗口的滑动方向。在滑动过程中，每次获得的时间窗口内的音素特征和声学特征为子音素特征和子声学特征，并且，在相同时间窗口内的子音素特征和子声学特征是对应于同一语音段的。

[98]在步骤404中，将所述子音素特征和对应的子声学特征输入至经训练的脸部关键点提取网络，得到与所述子音素特征和所述子声学特征对应的脸部关键点信息。如图6所示，对于每次获得的时间窗口内的子音素特征和子声学特征，所述脸部关键点提取网络输出该时间窗口对应的脸部关键点信息。

[99]示例性的，所述脸部关键点提取网络为3D脸部关键点提取网络，相应地，所得到的脸部关键点信息为3D脸部关键点信息。

[100]在步骤405中，获取所述3D脸部关键点信息对应的2D脸部关键点信息。

[101]在步骤406中，对多组2D脸部关键点信息进行滤波处理，使每个图像帧的2D脸部关键点信息与相邻帧的脸部关键点信息之间的变化量满足设定条件。

[102]在步骤407中，将经滤波处理后的每组2D脸部关键点信息与所述第二图像输入至脸部补全网络，得到与所述2D脸部关键点信息对应的目标脸部图像，其中，所述第二图像为遮挡脸部图像，第二图像中的下半脸被噪声填充以进行遮挡。

[103]在步骤408中，将步骤407中得到的多帧目标脸部图像（例如，说话人的脸部图像）与背景图像进行融合，得到第一图像序列。

[104]在步骤409中，根据所述第一图像序列与所述声音驱动数据对应的音频，得到所述目标对象的说话视频。

[105]图7是本公开至少一个实施例提出的说话视频生成装置的结构示意图；如图7所示，所述装置包括：第一获取单元701，用于获取声音驱动数据的音素特征以及声学特征，所述声音驱动数据包括音频、文本中的至少一项；第二获取单元702，用于根据所述音素特征和所述声学特征获取第一图像中目标对象的至少一组脸部关键点信息；第一得到

单元 703, 用于根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像, 得到与所述声音驱动数据对应的至少一个目标脸部图像, 其中, 所述第二图像中包括所述目标对象的特定部位的设定区域被遮挡; 第二得到单元 704, 用于根据所述声音驱动数据和所述至少一个目标脸部图像, 得到所述目标对象的说话视频。

[106]结合本公开提供的任一实施方式, 所述第一获取单元具体用于: 获取所述声音驱动数据对应的音频所包含的音素以及各个音素对应的时间戳, 得到所述声音驱动数据的音素特征; 对所述声音驱动数据对应的音频进行特征提取, 得到所述声音驱动数据的声学特征。

[107]结合本公开提供的任一实施方式, 所述第二获取单元具体用于: 获取所述音素特征所包含的多个子音素特征以及所述多个子音素特征对应的子声学特征; 将所述子音素特征和对应的子声学特征输入至脸部关键点提取网络, 得到与所述子音素特征和所述子声学特征对应的脸部关键点信息。

[108]结合本公开提供的任一实施方式, 所述脸部关键点信息包括 3D 脸部关键点信息, 所述装置还包括投影单元, 用于在根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像, 得到与所述声音驱动数据对应的至少一个目标脸部图像之前, 将所述 3D 脸部关键点信息投影到 2D 平面上, 得到所述 3D 脸部关键点信息对应的 2D 脸部关键点信息; 利用所述 2D 脸部关键点信息更新所述脸部关键点信息。

[109]结合本公开提供的任一实施方式, 所述装置还包括滤除单元, 用于在根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像, 得到与所述声音驱动数据对应的至少一个目标脸部图像之前, 对多组脸部关键点信息进行滤波处理, 使每个图像帧的脸部关键点信息与相邻帧的脸部关键点信息之间的变化量满足设定条件。

[110]结合本公开提供的任一实施方式, 所述第一得到单元具体用于: 将每组脸部关键点信息与所述第二图像输入至脸部补全网络, 得到与所述脸部关键点信息对应的目标脸部图像, 其中, 所述脸部补全网络用于根据脸部关键点信息对所述第二图像中被遮挡的设定区域进行补全。

[111]结合本公开提供的任一实施方式, 所述第二得到单元具体用于: 将所述至少一个目标脸部图像与设定背景图像进行融合, 得到第一图像序列; 根据所述第一图像序列与所述声音驱动数据对应的音频, 得到所述目标对象的说话视频。

[112]结合本公开提供的任一实施方式, 所述脸部关键点提取网络利用音素特征样本和对应的声学特征样本训练得到, 其中, 所述音素特征样本和所述声学特征样本包括标注的脸部关键点信息。

[113]结合本公开提供的任一实施方式, 所述脸部关键点提取网络通过以下方式训练得到: 根据所述音素特征样本和对应的声学特征样本, 对初始脸部关键点提取网络进行训练, 在网络损失的变化满足收敛条件时完成训练得到所述脸部关键点提取网络, 其中, 所述网络损失包括所述初始神经网络预测得到的脸部关键点信息与标注的脸部关键点信息之间的差异。

[114]结合本公开提供的任一实施方式, 所述音素特征样本和所述声学特征样本通过对一对象的音频的音素特征和声学特征进行所述对象的脸部关键点信息标注得到。

[115]结合本公开提供的任一实施方式, 所述音素特征样本和所述声学特征样本通过以下方式得到: 获取所述对象的说话视频; 根据所述说话视频获取多个脸部图像, 以及与每个所述脸部图像对应的至少一个音频帧; 获取每个所述脸部图像对应的至少一个音频帧的音素特征以及声学特征; 根据所述多个脸部图像获取脸部关键点信息, 并根据所述脸

部关键点信息对所述音素特征和所述声学特征进行标注,得到所述音素特征样本和所述声学特征样本。

[116]结合本公开提供的任一实施方式,所述脸部补全网络利用生成对抗网络训练得到,所述生成对抗网络包括所述脸部补全网络和第一鉴别网络,所述训练的网络损失包括:第一损失,用于指示所述脸部补全网络输出的脸部补全图像与完整脸部图像之间的差异,其中,所述完整脸部图像是所述脸部关键点信息对应的脸部图像;第二损失,用于指示所述第一鉴别网络对于输入图像输出的分类结果与所述输入图像的标注信息之间的差异,其中,所述标注信息指示所述输入图像为所述脸部补全网络输出的脸部补全图像或者为真实脸部图像。

[117]结合本公开提供的任一实施方式,所述生成对抗网络还包括第二鉴别网络,所述训练的网络损失还包括:第三损失,用于指示所述第二鉴别网络对于所述人脸补全图像与音素特征对应的判别结果与真实对应结果之间的差异。

[118]本公开至少一个实施例还提供了一种电子设备,如图8所示,所述设备包括存储器、处理器,存储器用于存储可在处理器上运行的计算机指令,处理器用于在执行所述计算机指令时实现本公开任一实施例所述的说话视频生成方法。

[119]本公开至少一个实施例还提供了一种计算机可读存储介质,其上存储有计算机程序,所述程序被处理器执行时实现本公开任一实施例所述的说话视频生成方法。

[120]本领域技术人员应明白,本说明书一个或多个实施例可提供为方法、系统或计算机程序产品。因此,本说明书一个或多个实施例可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且,本说明书一个或多个实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[121]本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于数据处理设备实施例而言,由于其基本相似于方法实施例,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。

[122]上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下,在权利要求书中记载的行为或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外,在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中,多任务处理和并行处理也是可以的或者可能是有利的。

[123]本说明书中描述的主题及功能操作的实施例可以在以下中实现:数字电子电路、有形体现的计算机软件或固件、包括本说明书中公开的结构及其结构性等同物的计算机硬件、或者它们中的一个或多个的组合。本说明书中描述的主题的实施例可以实现为一个或多个计算机程序,即编码在有形非暂时性程序载体上以被数据处理装置执行或控制数据处理装置的操作的计算机程序指令中的一个或多个模块。可替代地或附加地,程序指令可以被编码在人工生成的传播信号上,例如机器生成的电、光或电磁信号,该信号被生成以将信息编码并传输到合适的接收机装置以由数据处理装置执行。计算机存储介质可以是机器可读存储设备、机器可读存储基板、随机或串行存取存储器设备、或它们中的一个或多个的组合。

[124]本说明书中描述的处理及逻辑流程可以由执行一个或多个计算机程序的一个或多个可编程计算机执行,以通过根据输入数据进行操作并生成输出来执行相应的功能。所述处理及逻辑流程还可以由专用逻辑电路—例如 FPGA(现场可编程门阵列)或 ASIC

(专用集成电路)来执行,并且装置也可以实现为专用逻辑电路。

[125]适合用于执行计算机程序的计算机包括,例如通用和/或专用微处理器,或任何其他类型的中央处理单元。通常,中央处理单元将从只读存储器和/或随机存取存储器接收指令和数据。计算机的基本组件包括用于实施或执行指令的中央处理单元以及用于存储指令和数据的一个或多个存储器设备。通常,计算机还将包括用于存储数据的一个或多个大容量存储设备,例如磁盘、磁光盘或光盘等,或者计算机将可操作地与此大容量存储设备耦接以从其接收数据或向其传送数据,抑或两种情况兼而有之。然而,计算机不是必须具有这样的设备。此外,计算机可以嵌入在另一设备中,例如移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏操纵台、全球定位系统(GPS)接收机、或例如通用串行总线(USB)闪存驱动器的便携式存储设备,仅举几例。

[126]适合于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、媒介和存储器设备,例如包括半导体存储器设备(例如 EPROM、EEPROM 和闪存设备)、磁盘(例如内部硬盘或可移动盘)、磁光盘以及 CD ROM 和 DVD-ROM 盘。处理器和存储器可由专用逻辑电路补充或并入专用逻辑电路中。

[127]虽然本说明书包含许多具体实施细节,但是这些不应被解释为限制任何发明的范围或所要求保护的发明,而是主要用于描述特定发明的具体实施例的特征。本说明书内在多个实施例中描述的某些特征也可以在单个实施例中被组合实施。另一方面,在单个实施例中描述的各种特征也可以在多个实施例中分开实施或以任何合适的子组合来实施。此外,虽然特征可以如上所述在某些组合中起作用并且甚至最初如此要求保护,但是来自所要求保护的组合中的一个或多个特征在一些情况下可以从该组合中去除,并且所要求保护的组合可以指向子组合或子组合的变型。

[128]类似地,虽然在附图中以特定顺序描绘了操作,但是这不应被理解为要求这些操作以所示的特定顺序执行或顺次执行、或者要求所有例示的操作被执行,以实现期望的结果。在某些情况下,多任务和并行处理可能是有利的。此外,上述实施例中的各种系统模块和组件的分离不应被理解为在所有实施例中均需要这样的分离,并且应当理解,所描述的程序组件和系统通常可以一起集成在单个软件产品中,或者封装成多个软件产品。

[129]由此,主题的特定实施例已被描述。其他实施例在所附权利要求书的范围以内。在某些情况下,权利要求书中记载的动作可以以不同的顺序执行并且仍实现期望的结果。此外,附图中描绘的处理并非必需所示的特定顺序或顺次顺序,以实现期望的结果。在某些实现中,多任务和并行处理可能是有利的。

[130]以上所述仅为本说明书一个或多个实施例的较佳实施例而已,并不用以限制本说明书一个或多个实施例,凡在本说明书一个或多个实施例的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本说明书一个或多个实施例保护的范围之内。

权利要求书

1、一种说话视频生成方法，其特征在于，所述方法包括：

获取声音驱动数据的音素特征以及声学特征，所述声音驱动数据包括音频、文本中的至少一项；

根据所述音素特征和所述声学特征获取第一图像中目标对象的至少一组脸部关键点信息；

根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像，其中，所述第二图像中包括所述目标对象的特定部位的设定区域被遮挡；

根据所述声音驱动数据和所述至少一个目标脸部图像，得到所述目标对象的说话视频。

2、根据权利要求1所述的方法，其特征在于，所述获取声音驱动数据的音素特征以及声学特征，包括：

获取所述声音驱动数据对应的音频所包含的音素以及各个音素对应的时间戳，得到所述声音驱动数据的音素特征；

对所述声音驱动数据对应的音频进行特征提取，得到所述声音驱动数据的声学特征。

3、根据权利要求1或2所述的方法，其特征在于，所述根据所述音素特征和所述声学特征获取第一图像中目标对象的至少一组脸部关键点信息，包括：

获取所述音素特征所包含的多个子音素特征以及所述多个子音素特征对应的子声学特征；

将所述子音素特征和对应的子声学特征输入至脸部关键点提取网络，得到与所述子音素特征和所述子声学特征对应的脸部关键点信息。

4、根据权利要求1至3任一项所述的方法，其特征在于，所述脸部关键点信息包括3维（3D）脸部关键点信息，在根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像之前，所述方法还包括：

将所述3D脸部关键点信息投影到2维（2D）平面上，得到所述3D脸部关键点信息对应的2D脸部关键点信息；

利用所述2D脸部关键点信息更新所述脸部关键点信息。

5、根据权利要求1至4任一项所述的方法，其特征在于，在根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像之前，所述方法还包括：

对多组脸部关键点信息进行滤波处理，使每个图像帧的脸部关键点信息与该图像帧的相邻帧的脸部关键点信息之间的变化量满足设定条件。

6、根据权利要求1至5任一项所述的方法，其特征在于，所述根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像，得到与所述声音驱动数据对应的至少一个目标脸部图像，包括：

将每组脸部关键点信息与所述第二图像输入至脸部补全网络，得到与所述脸部关键点信息对应的目标脸部图像，其中，所述脸部补全网络用于根据脸部关键点信息对所述第二图像中被遮挡的设定区域进行补全。

7、根据权利要求1至6任一项所述的方法，其特征在于，所述根据所述声音驱动数据和所述至少一个目标脸部图像，得到所述目标对象的说话视频，包括：

将所述至少一个目标脸部图像与设定背景图像进行融合，得到第一图像序列；

根据所述第一图像序列与所述声音驱动数据对应的音频，得到所述目标对象的说话视频。

8、根据权利要求3至7任一项所述的方法，其特征在于，所述脸部关键点提取网

络利用音素特征样本和对应的声学特征样本训练得到,其中,所述音素特征样本和所述声学特征样本包括标注的脸部关键点信息。

9、根据权利要求8所述的方法,其特征在于,所述脸部关键点提取网络通过以下方式训练得到:

根据所述音素特征样本和对应的声学特征样本,对初始脸部关键点提取网络进行训练,在网络损失的变化满足收敛条件时完成训练得到所述脸部关键点提取网络,其中,所述网络损失包括所述初始神经网络预测得到的脸部关键点信息与标注的脸部关键点信息之间的差异。

10、根据权利要求8或9所述的方法,其特征在于,所述音素特征样本和所述声学特征样本通过对一对象的音频的音素特征和声学特征进行所述对象的脸部关键点信息标注得到。

11、根据权利要求10所述的方法,其特征在于,所述音素特征样本和所述声学特征样本通过以下方式得到:

获取所述对象的说话视频;

根据所述说话视频获取多个脸部图像,以及与每个所述脸部图像对应的至少一个音频帧;

获取每个所述脸部图像对应的至少一个音频帧的音素特征以及声学特征;

根据所述多个脸部图像获取脸部关键点信息,并根据所述脸部关键点信息对所述音素特征和所述声学特征进行标注,得到所述音素特征样本和所述声学特征样本。

12、根据权利要求6至11任一项所述的方法,其特征在于,所述脸部补全网络利用生成对抗网络训练得到,所述生成对抗网络包括所述脸部补全网络和第一鉴别网络,所述训练的网络损失包括:

第一损失,用于指示所述脸部补全网络输出的脸部补全图像与完整脸部图像之间的差异,其中,所述完整脸部图像是所述脸部关键点信息对应的脸部图像;

第二损失,用于指示所述第一鉴别网络对于输入图像输出的分类结果与所述输入图像的标注信息之间的差异,其中,所述标注信息指示所述输入图像为所述脸部补全网络输出的脸部补全图像或者为真实脸部图像。

13、根据权利要求12所述的方法,其特征在于,所述生成对抗网络还包括第二鉴别网络,所述训练的网络损失还包括:

第三损失,用于指示所述第二鉴别网络对于所述脸部补全图像与音素特征之间的对应的判别结果与真实对应结果之间的差异。

14、一种说话视频生成装置,其特征在于,所述装置包括:

第一获取单元,用于获取声音驱动数据的音素特征以及声学特征,所述声音驱动数据包括音频、文本中的至少一项;

第二获取单元,用于根据所述音素特征和所述声学特征获取第一图像中目标对象的至少一组脸部关键点信息;

第一得到单元,用于根据所述至少一组脸部关键点信息以及包含所述目标对象的脸部的第二图像,得到与所述声音驱动数据对应的至少一个目标脸部图像,其中,所述第二图像中包括所述目标对象的特定部位的设定区域被遮挡;

第二得到单元,用于根据所述声音驱动数据和所述至少一个目标脸部图像,得到所述目标对象的说话视频。

15、一种电子设备,其特征在于,所述设备包括存储器、处理器,所述存储器用于存储可在处理器上运行的计算机指令,所述处理器用于在执行所述计算机指令时实现权利要求1至13任一项所述的方法。

16、一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述程序被处理器执行时实现权利要求1至13任一所述的方法。

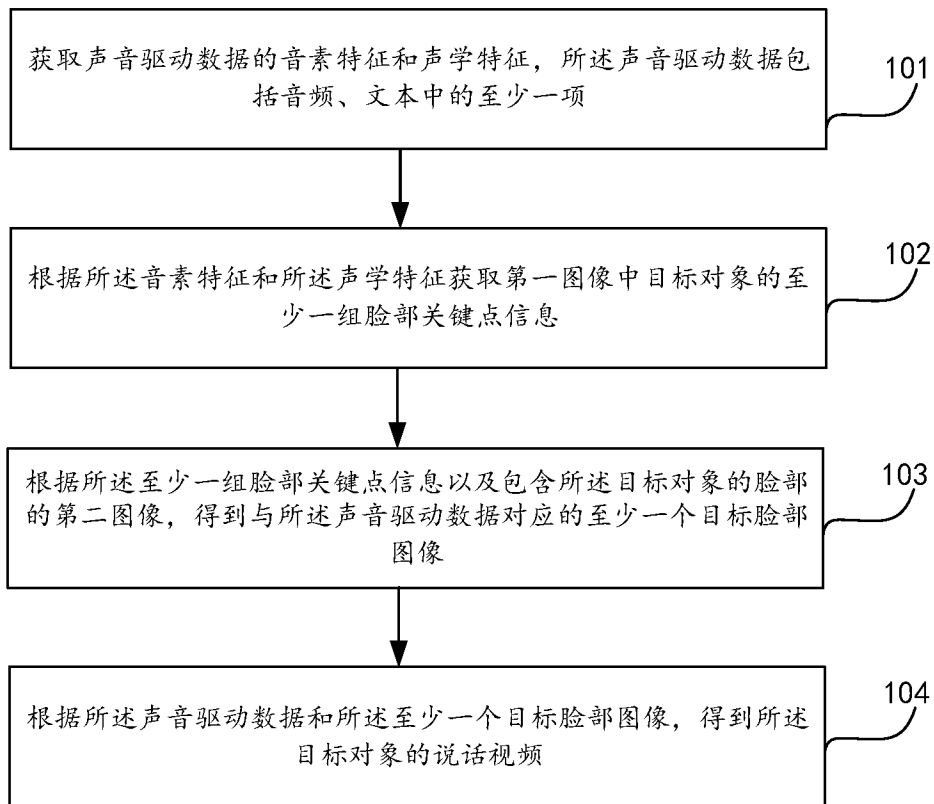


图 1

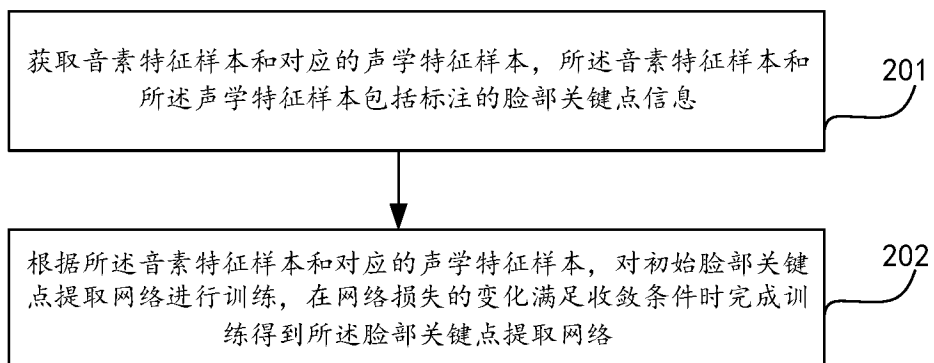


图 2

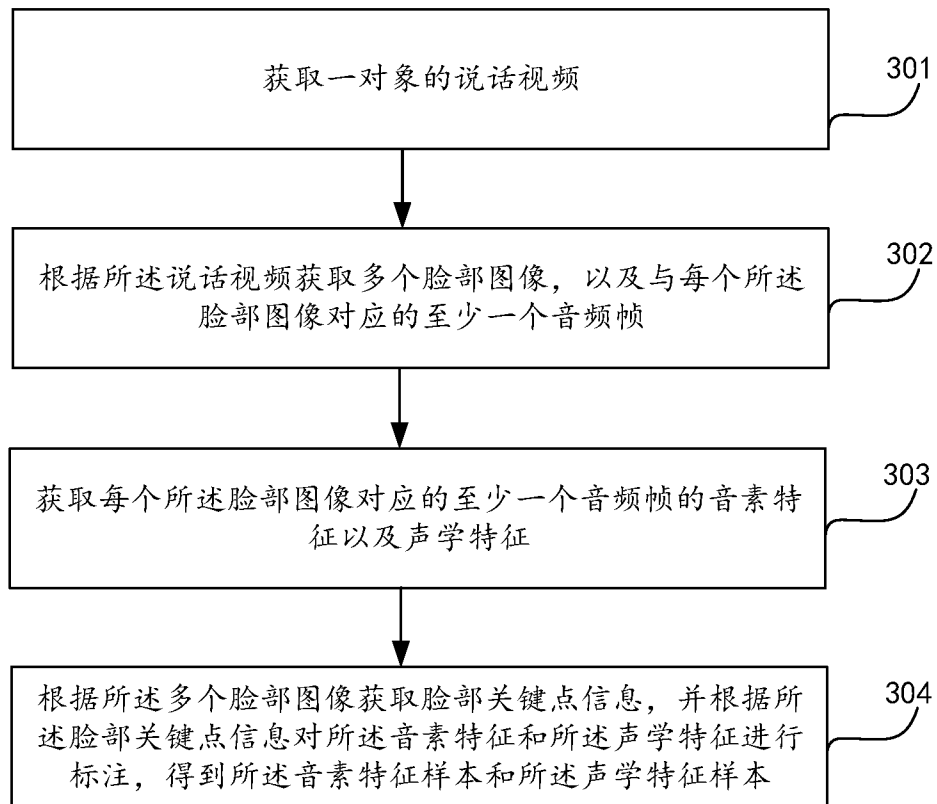


图 3

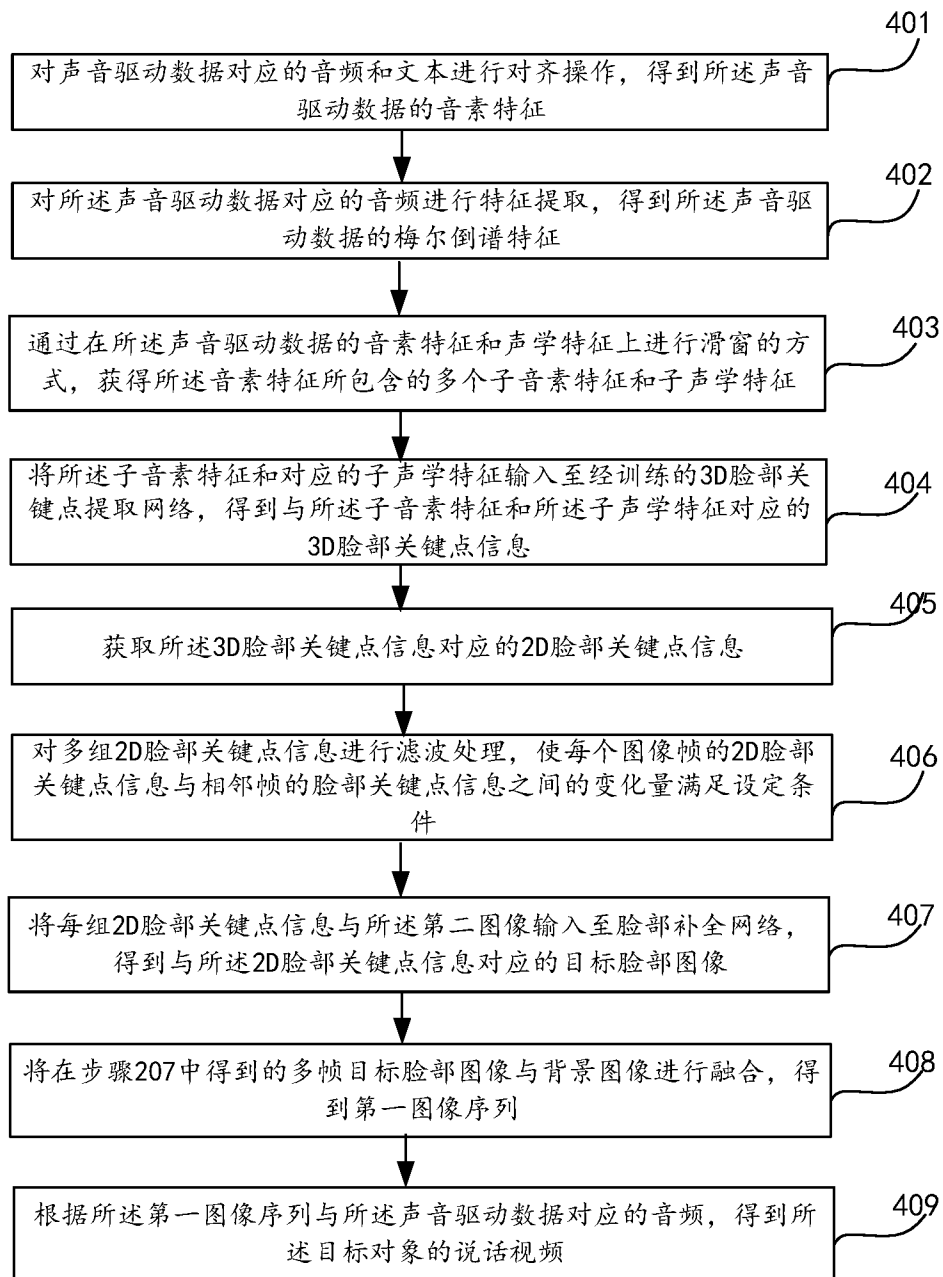


图 4

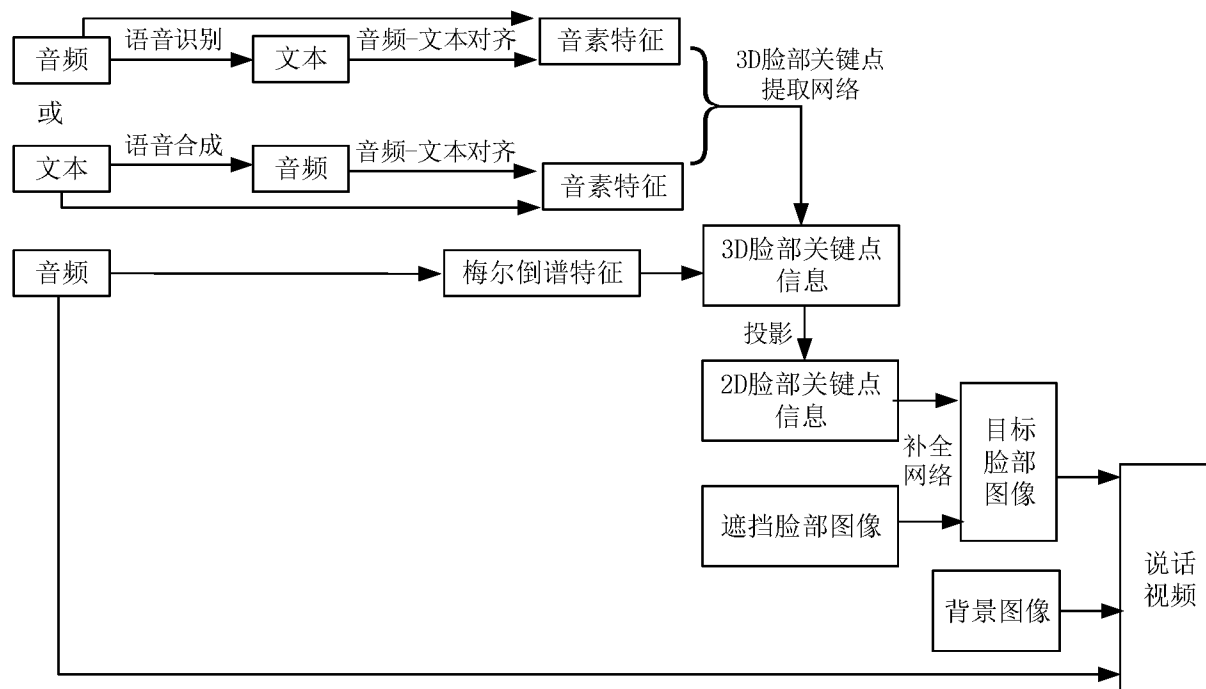


图 5

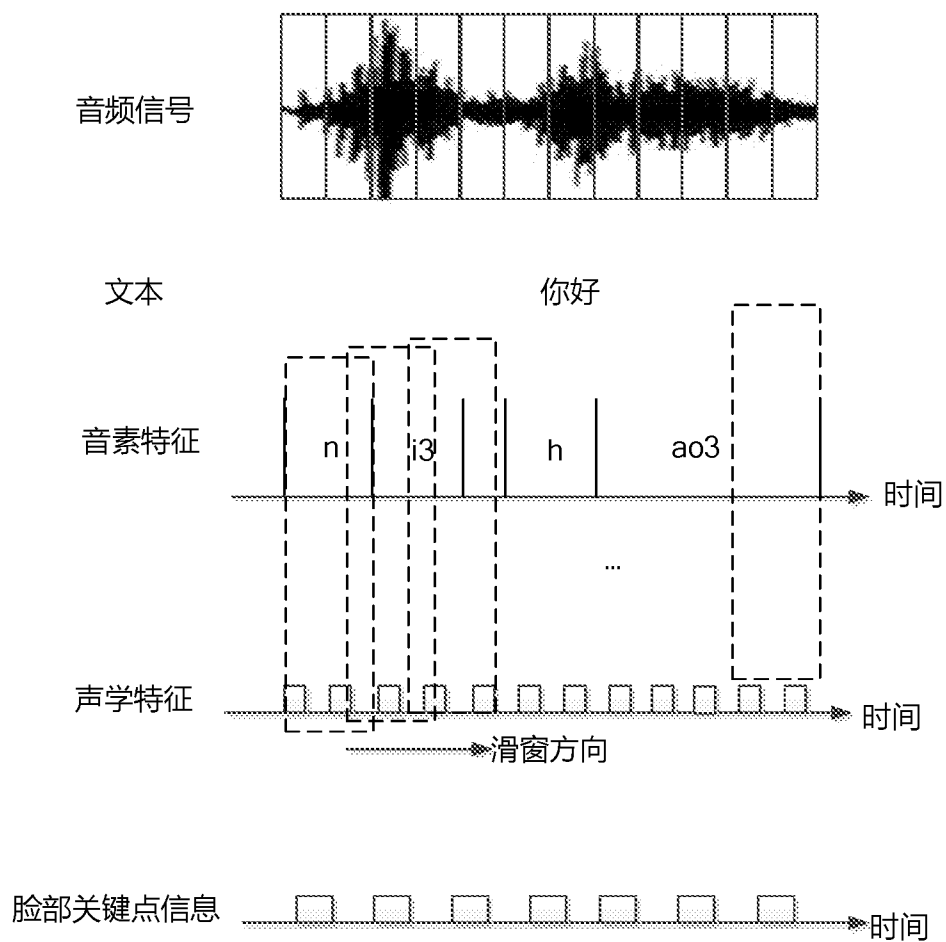


图 6

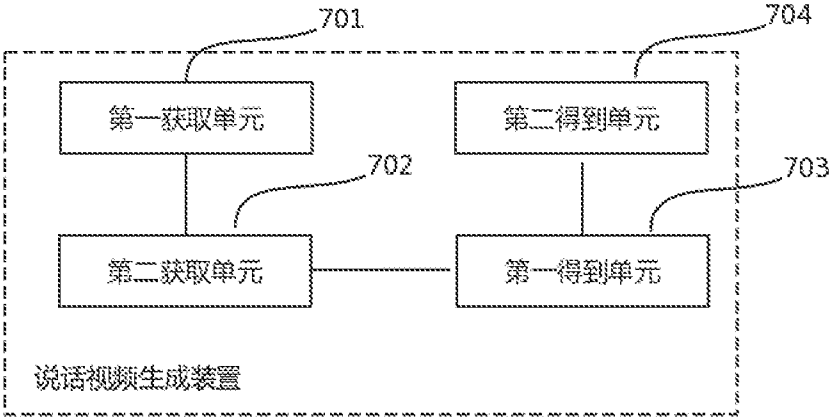


图 7

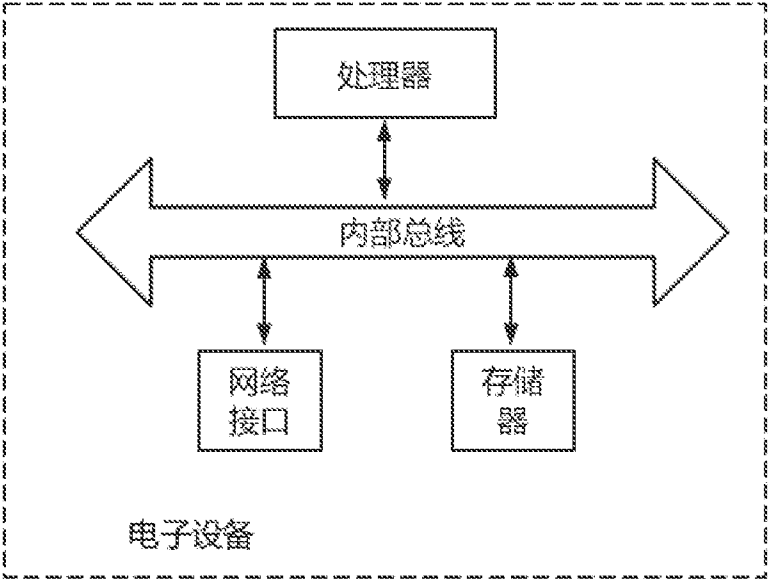


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/128584

A. CLASSIFICATION OF SUBJECT MATTER

G10L 15/26(2006.01)i; G06N 3/04(2023.01)i; H04N 21/234(2011.01)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)

G10L15/-; G06N3/-; H04N21/-

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)

CNABS, WPABS, CNTXT, VCN, VEN, DWPI, WOTXT, EPTXT, USTXT: 商汤科技, 王宇欣, 吴文岩, 说话, 讲话, 对话, 声音, 音频, 语音, 驱动, 支配, 文本, 图像, 脸部, 面部, 嘴, 视频, speak, talk, speech, voice, sound, driv+, text, imag+, pictur+, face, mouth, video

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 114093384 A (SHANGHAI SENSETIME TECHNOLOGY DEVELOPMENT CO., LTD.) 25 February 2022 (2022-02-25) claims 1-16	1-16
A	CN 113228163 A (SNAP INC.) 06 August 2021 (2021-08-06) paragraphs [0037]-[0041], and figures 3 and 4	1-16
A	CN 111432233 A (BEIJING BYTEDANCE NETWORK TECHNOLOGY CO., LTD.) 17 July 2020 (2020-07-17) entire document	1-16
A	US 2021312915 A1 (HI AUTO LTD.) 07 October 2021 (2021-10-07) entire document	1-16
A	JP 3633399 B2 (MITSUBISHI ELECTRIC INFORMATION TECHNOLOGY CENTER AMERICA, INC.) 30 March 2005 (2005-03-30) entire document	1-16
A	JP 2000123192 A (MITSUBISHI ELECTRIC INFORMATION TECHNOLOGY CENTER AMERICA, INC.) 28 April 2000 (2000-04-28) 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

13 January 2023

Date of mailing of the international search report

20 January 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/128584

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 112215926 A (BEIJING HUAYAN HUYU TECHNOLOGY CO., LTD.) 12 January 2021 (2021-01-12) entire document	1-16
A	CN 110677598 A (BEIJING SENSETIME SCIENCE TECHNOLOGY DEVELOPMENT CO., LTD.) 10 January 2020 (2020-01-10) entire document	1-16
A	CN 109308731 A (ZHEJIANG UNIVERSITY) 05 February 2019 (2019-02-05) entire document	1-16
A	CN 113542624 A (ALIBABA SINGAPORE HOLDINGS LIMITED) 22 October 2021 (2021-10-22) entire document	1-16
A	EP 0992933 A3 (MITSUBISHI ELECTRIC CORP.) 14 August 2002 (2002-08-14) entire document	1-16
A	CN 112562722 A (XINHUA ZHIYUN TECHNOLOGY CO., LTD.) 26 March 2021 (2021-03-26) entire document	1-16
A	EP 0992933 A2 (MITSUBISHI ELECTRIC CORP.) 12 April 2000 (2000-04-12) entire document	1-16
A	CN 112667068 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 16 April 2021 (2021-04-16) entire document	1-16
A	CN 110555507 A (SHENZHEN ZHUIYI TECHNOLOGY CO., LTD.) 10 December 2019 (2019-12-10) entire document	1-16
A	US 6735566 B1 (MITSUBISHI ELECTRIC RESEARCH LABORATORIES, INC.) 11 May 2004 (2004-05-11) entire document	1-16
A	CN 111741326 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 02 October 2020 (2020-10-02) entire document	1-16

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/128584

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	114093384	A	25 February 2022	None			
CN	113228163	A	06 August 2021	KR	20210119441	A	05 October 2021
				WO	2020150688	A1	23 July 2020
				EP	3912159	A1	24 November 2021
CN	111432233	A	17 July 2020	None			
US	2021312915	A1	07 October 2021	None			
JP	3633399	B2	30 March 2005	EP	0992933	A2	12 April 2000
				JP	2000123192	A	28 April 2000
				US	6735566	B1	11 May 2004
JP	2000123192	A	28 April 2000	EP	0992933	A2	12 April 2000
				US	6735566	B1	11 May 2004
CN	112215926	A	12 January 2021	None			
CN	110677598	A	10 January 2020	WO	2021052224	A1	25 March 2021
				KR	20210140762	A	23 November 2021
				JP	2022526148	A	23 May 2022
				SG	11202108498R	A	29 September 2021
				US	2021357625	A1	18 November 2021
CN	109308731	A	05 February 2019	None			
CN	113542624	A	22 October 2021	None			
EP	0992933	A3	14 August 2002	EP	0992933	A2	12 April 2000
				JP	2000123192	A	28 April 2000
				US	6735566	B1	11 May 2004
CN	112562722	A	26 March 2021	None			
EP	0992933	A2	12 April 2000	JP	2000123192	A	28 April 2000
				US	6735566	B1	11 May 2004
CN	112667068	A	16 April 2021	None			
CN	110555507	A	10 December 2019	None			
US	6735566	B1	11 May 2004	EP	0992933	A2	12 April 2000
				JP	2000123192	A	28 April 2000
CN	111741326	A	02 October 2020	None			

国际检索报告

国际申请号

PCT/CN2022/128584

A. 主题的分类 G10L 15/26(2006.01)i; G06N 3/04(2023.01)i; H04N 21/234(2011.01)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) G10L15/-; G06N3/-; H04N21/- 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)) CNABS, WPABS, CNTXT, VCN, VEN, DWPI, WOTXT, EPTXT, USTXT; 商汤科技, 王宇欣, 吴文岩, 说话, 讲话, 对话, 声音, 音频, 语音, 驱动, 支配, 文本, 图像, 脸部, 面部, 嘴, 视频, speak, talk, speech, voice, sound, driv+, text, imag+, pictur+, face, mouth, video				
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 114093384 A (上海商汤科技开发有限公司) 2022年2月25日 (2022 - 02 - 25) 权利要求1-16	1-16		
A	CN 113228163 A (斯纳普公司) 2021年8月6日 (2021 - 08 - 06) 第[0037]-[0041]段, 附图3, 4	1-16		
A	CN 111432233 A (北京字节跳动网络技术有限公司) 2020年7月17日 (2020 - 07 - 17) 全文	1-16		
A	US 2021312915 A1 (HI AUTO LTD.) 2021年10月7日 (2021 - 10 - 07) 全文	1-16		
A	JP 3633399 B2 (MITSUBISHI ELECTRIC INF. TECH.) 2005年3月30日 (2005 - 03 - 30) 全文	1-16		
A	JP 2000123192 A (MITSUBISHI ELECTRIC INF. TECH.) 2000年4月28日 (2000 - 04 - 28) 全文	1-16		
A	CN 112215926 A (北京华严互娱科技有限公司) 2021年1月12日 (2021 - 01 - 12) 全文	1-16		
☒ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。				
<table><tr><td>* 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</td><td>“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</td></tr></table>			* 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
* 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 2023年1月13日		国际检索报告邮寄日期 2023年1月20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 董小东 电话号码 (86-28) 62967922		

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
A	CN 110677598 A (北京市商汤科技开发有限公司) 2020年1月10日 (2020 - 01 - 10) 全文	1-16
A	CN 109308731 A (浙江大学) 2019年2月5日 (2019 - 02 - 05) 全文	1-16
A	CN 113542624 A (阿里巴巴新加坡控股有限公司) 2021年10月22日 (2021 - 10 - 22) 全文	1-16
A	EP 0992933 A3 (MITSUBISHI ELECTRIC CORP.) 2002年8月14日 (2002 - 08 - 14) 全文	1-16
A	CN 112562722 A (新华智云科技有限公司) 2021年3月26日 (2021 - 03 - 26) 全文	1-16
A	EP 0992933 A2 (MITSUBISHI ELECTRIC CORP.) 2000年4月12日 (2000 - 04 - 12) 全文	1-16
A	CN 112667068 A (北京百度网讯科技有限公司) 2021年4月16日 (2021 - 04 - 16) 全文	1-16
A	CN 110555507 A (深圳追一科技有限公司) 2019年12月10日 (2019 - 12 - 10) 全文	1-16
A	US 6735566 B1 (MITSUBISHI ELECTRIC RES LAB.) 2004年5月11日 (2004 - 05 - 11) 全文	1-16
A	CN 111741326 A (腾讯科技深圳有限公司) 2020年10月2日 (2020 - 10 - 02) 全文	1-16

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/128584

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	114093384	A	2022年2月25日	无			
CN	113228163	A	2021年8月6日	KR	20210119441	A	2021年10月5日
				WO	2020150688	A1	2020年7月23日
				EP	3912159	A1	2021年11月24日
CN	111432233	A	2020年7月17日	无			
US	2021312915	A1	2021年10月7日	无			
JP	3633399	B2	2005年3月30日	EP	0992933	A2	2000年4月12日
				JP	2000123192	A	2000年4月28日
				US	6735566	B1	2004年5月11日
JP	2000123192	A	2000年4月28日	EP	0992933	A2	2000年4月12日
				US	6735566	B1	2004年5月11日
CN	112215926	A	2021年1月12日	无			
CN	110677598	A	2020年1月10日	WO	2021052224	A1	2021年3月25日
				KR	20210140762	A	2021年11月23日
				JP	2022526148	A	2022年5月23日
				SG	11202108498R	A	2021年9月29日
				US	2021357625	A1	2021年11月18日
CN	109308731	A	2019年2月5日	无			
CN	113542624	A	2021年10月22日	无			
EP	0992933	A3	2002年8月14日	EP	0992933	A2	2000年4月12日
				JP	2000123192	A	2000年4月28日
				US	6735566	B1	2004年5月11日
CN	112562722	A	2021年3月26日	无			
EP	0992933	A2	2000年4月12日	JP	2000123192	A	2000年4月28日
				US	6735566	B1	2004年5月11日
CN	112667068	A	2021年4月16日	无			
CN	110555507	A	2019年12月10日	无			
US	6735566	B1	2004年5月11日	EP	0992933	A2	2000年4月12日
				JP	2000123192	A	2000年4月28日
CN	111741326	A	2020年10月2日	无			