(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110189741 B
(45)授权公告日 2024.09.06

(21)申请号 201810730283.3
(22)申请日 2018.07.05
(65)同一申请的已公布的文献号
申请公布号 CN 110189741 A
(43)申请公布日 2019.08.30
(73)专利权人 腾讯数码(天津)有限公司
地址 300000 天津市滨海新区经济技术开发区第四大街80号天大科技园软件大厦北楼2层202室
(72)发明人 崔凌睿 芦艺 周怡婷 吴歆婉 梁艺东 梅晓 冯启航 王方晓 江会福 郑尚镇 余乐 夏胜飞 王敬轩 张然 郭艺帆 张震云
(74)专利代理机构 华进联合专利商标代理有限公司 44224
专利代理师 李文渊 邓云鹏
(51)Int.Cl.
G10L 13/04(2013.01)
(56)对比文件
JP 2002132281 A,2002.05.09
审查员 王雨辰

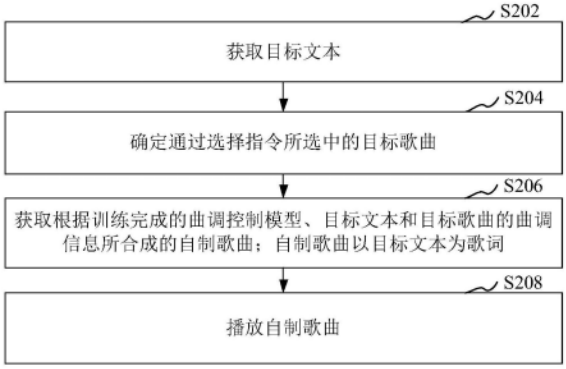
权利要求书4页 说明书15页 附图5页

(54)发明名称

音频合成方法、装置、存储介质和计算机设备

(57)摘要

本发明涉及一种音频合成方法、装置、存储介质和计算机设备，该方法包括：获取目标文本；确定通过选择指令所选中的目标歌曲；获取根据训练完成的曲调控制模型、所述目标文本和所述目标歌曲的曲调信息所合成的自制歌曲；所述自制歌曲以所述目标文本为歌词；播放所述自制歌曲。本申请提供的方案提高了音频播放的效果。



CN 110189741 B

1.一种音频合成方法,包括:

响应于用户指令展示视频录制应用的背景音频配置界面;在所述背景音频配置界面,展示文本输入框、文本模板列表和候选歌曲;获取通过所述文本输入框输入的目标文本,或者获取从所述文本模板列表中选中的目标文本;确定通过选择指令从所述候选歌曲中所选中的目标歌曲;

接收选择指令;

当所述选择指令选中目标歌曲时,将所述目标文本和所述目标歌曲的曲调信息共同输入训练完成的曲调控制模型;通过所述曲调控制模型的隐层对所述目标文本中的各字符进行曲调匹配,确定各字符各自相应的曲调特征,通过曲调控制模型的输出层,输出将所述目标文本中各字符按照相应的曲调特征进行语音合成得到的自制歌曲;所述自制歌曲以所述目标文本为歌词;

播放所述自制歌曲;

当所述选择指令选中目标说话对象时,则获取根据训练完成的音色控制模型和所述目标文本所合成的自制音频;所述音色控制模型与所述目标说话对象匹配;

将所述自制歌曲或所述自制音频配置为背景音频;展示视频录制页面;响应于用户指令播放所述背景音频并采集图像录制视频。

2.根据权利要求1所述的方法,其特征在于,所述方法还包括:

将所述目标文本和所述目标歌曲的歌曲标识发送至服务器;

通过所述服务器查找与所述歌曲标识对应的曲调信息;

所述将所述目标文本和所述目标歌曲的曲调信息共同输入训练完成的曲调控制模型;通过所述曲调控制模型的隐层对所述目标文本中的各字符进行曲调匹配,确定各字符各自相应的曲调特征,通过曲调控制模型的输出层,输出将所述目标文本中各字符按照相应的曲调特征进行语音合成得到的自制歌曲,包括:

通过所述服务器将所述目标文本和所述曲调信息共同输入训练完成的曲调控制模型;通过所述曲调控制模型的隐层对所述目标文本中的各字符进行曲调匹配,确定各字符各自相应的曲调特征,通过曲调控制模型的输出层,输出将所述目标文本中各字符按照相应的曲调特征进行语音合成得到的自制歌曲;接收所述服务器所反馈的自制歌曲。

3.根据权利要求1所述的方法,其特征在于,所述方法还包括训练曲调控制模型的步骤;所述目标歌曲选自候选歌曲;

所述训练曲调控制模型,包括:

收集候选歌曲所对应的候选歌曲音频;

根据收集的候选歌曲音频,确定各候选歌曲所对应的候选歌曲曲调;

获取文本样本;

根据所述文本样本和所述候选歌曲曲调训练得到所述曲调控制模型。

4.根据权利要求1所述的方法,其特征在于,所述方法还包括训练与各候选说话对象匹配的音色控制模型的步骤;所述目标说话对象选自候选说话对象;

所述训练与各候选说话对象匹配的音色控制模型,包括:

收集各候选说话对象相应的音频素材;

根据各音频素材确定相应候选说话对象所相应的音素素材序列;

采用各候选说话对象所相应的音素素材序列,训练得到各候选说话对象匹配的音色控制模型。

5.根据权利要求1所述的方法,其特征在于,所述获取根据训练完成的音色控制模型和所述目标文本所合成的自制音频,包括:

查询与所述目标说话对象匹配的音色控制模型;

确定所述目标文本所对应的音素序列;

通过所述音色控制模型根据所述音素序列合成自制语音;

根据所述自制语音和背景伴奏合成自制音频。

6.根据权利要求1所述的方法,其特征在于,所述接收选择指令,包括:

接收对应于虚拟对象附加元素的选择指令;

确定所述选择指令选中的虚拟对象附加元素所对应的目标说话对象;

所述方法还包括:

将所述自制音频配置为背景音频;

将所述虚拟对象附加元素叠加至采集的图像得到视频帧;

基于所述背景音频和叠加得到的所述视频帧生成录制的视频。

7.根据权利要求1所述的方法,其特征在于,所述接收选择指令,包括:

接收用于进行模拟视频通话的选择指令;

确定所述选择指令选中的目标说话对象所对应的图片;

所述方法还包括:

将所述自制音频配置为背景音频;

根据所述图片和采集的图像生成通话视频帧;

基于所述背景音频和生成的所述通话视频帧生成录制的视频。

8.一种音频合成装置,包括:

获取模块,用于响应于用户指令展示视频录制应用的背景音频配置界面;在所述背景音频配置界面,展示文本输入框、文本模板列表和候选歌曲;获取通过所述文本输入框输入的目标文本,或者获取从所述文本模板列表中选中的目标文本;

确定模块,用于确定通过选择指令从所述候选歌曲中所选中的目标歌曲;

获取模块,还用于接收选择指令;当所述选择指令选中目标歌曲时,将所述目标文本和所述目标歌曲的曲调信息共同输入训练完成的曲调控制模型;通过所述曲调控制模型的隐层对所述目标文本中的各字符进行曲调匹配,确定各字符各自相应的曲调特征,通过曲调控制模型的输出层,输出将所述目标文本中各字符按照相应的曲调特征进行语音合成得到的自制歌曲;所述自制歌曲以所述目标文本为歌词;

播放模块,用于播放所述自制歌曲;

所述获取模块,还用于当所述选择指令选中目标说话对象时,则获取根据训练完成的音色控制模型和所述目标文本所合成的自制音频;所述音色控制模型与所述目标说话对象匹配;

录制模块,用于将所述自制歌曲或所述自制音频配置为背景音频;展示视频录制页面;响应于用户指令播放所述背景音频并采集图像录制视频。

9.根据权利要求8所述的装置,其特征在于,所述获取模块还用于:

将所述目标文本和所述目标歌曲的歌曲标识发送至服务器;通过所述服务器查找到与所述歌曲标识对应的曲调信息;通过所述服务器将所述目标文本和所述曲调信息共同输入训练完成的曲调控制模型;通过所述曲调控制模型的隐层对所述目标文本中的各字符进行曲调匹配,确定各字符各自相应的曲调特征,通过曲调控制模型的输出层,输出将所述目标文本中各字符按照相应的曲调特征进行语音合成得到的自制歌曲;接收所述服务器所反馈的自制歌曲。

10.根据权利要求8所述的装置,其特征在于,所述装置还包括训练模块,用于:训练曲调控制模型;

所述训练模块还用于:

收集候选歌曲所对应的候选歌曲音频;

根据收集的候选歌曲音频,确定各候选歌曲所对应的候选歌曲曲调;

获取文本样本;

根据所述文本样本和所述候选歌曲曲调训练得到所述曲调控制模型;其中,所述目标歌曲选自候选歌曲。

11.根据权利要求8所述的装置,其特征在于,所述装置还包括训练模块,用于训练与各候选说话对象匹配的音色控制模型;

所述训练模块还用于:

收集各候选说话对象相应的音频素材;

根据各音频素材确定相应候选说话对象所相应的音素素材序列;

采用各候选说话对象所相应的音素素材序列,训练得到各候选说话对象匹配的音色控制模型,其中,所述目标说话对象选自候选说话对象。

12.根据权利要求8所述的装置,其特征在于,所述获取模块还用于:

查询与所述目标说话对象匹配的音色控制模型;确定所述目标文本所对应的音素序列;通过所述音色控制模型根据所述音素序列合成自制语音;根据所述自制语音和背景伴奏合成自制音频。

13.根据权利要求8所述的装置,其特征在于,所述获取模块,还用于:

接收对应于虚拟对象附加元素的选择指令;

确定所述选择指令选中的虚拟对象附加元素所对应的目标说话对象;

所述录制模块还用于:

将所述自制音频配置为背景音频;

将所述虚拟对象附加元素叠加至采集的图像得到视频帧;

基于所述背景音频和叠加得到的所述视频帧生成录制的视频。

14.根据权利要求8所述的装置,其特征在于,所述获取模块还用于:

接收用于进行模拟视频通话的选择指令;

确定所述选择指令选中的目标说话对象所对应的图片;

所述录制模块还用于:

将所述自制音频配置为背景音频;

根据所述图片和采集的图像生成通话视频帧;

基于所述背景音频和生成的所述通话视频帧生成录制的视频。

15.一种计算机可读存储介质,所述计算机可读存储介质上存储有计算机程序,所述计算机程序被处理器执行时,使得所述处理器执行如权利要求1至7中任一项所述的方法的步骤。

16.一种计算机设备,包括存储器和处理器,所述存储器中储存有计算机程序,所述计算机程序被所述处理器执行时,使得所述处理器执行如权利要求1至7中任一项所述的方法的步骤。

音频合成方法、装置、存储介质和计算机设备

技术领域

[0001] 本发明涉及计算机技术领域,特别是涉及一种音频合成方法、装置、存储介质和计算机设备。

背景技术

[0002] 随着计算机技术的发展,越来越多的计算机设备都具有音频播放功能,比如智能手机以及平板电脑等。这样用户为了缓解紧张的工作情绪并增加生活乐趣,越来越普遍地选择在闲暇时聆听歌曲,或者使用歌曲来进行视频拍摄等。

[0003] 目前,用户通常会从歌曲库中选择歌曲来播放。但是,由于版权等问题导致歌曲库中可选的歌曲数量有限,用户难以从这些歌曲中选中符合自己期望的歌曲来播放,从而导致音频播放的效果较差。

发明内容

[0004] 基于此,有必要针对目前音频播放的效果较差的问题,提供一种音频合成方法、装置、存储介质和计算机设备。

[0005] 一种音频合成方法,包括:

[0006] 获取目标文本;

[0007] 确定通过选择指令所选中的目标歌曲;

[0008] 获取根据训练完成的曲调控制模型、所述目标文本和所述目标歌曲的曲调信息所合成的自制歌曲;所述自制歌曲以所述目标文本为歌词;

[0009] 播放所述自制歌曲。

[0010] 一种音频合成装置,包括:

[0011] 获取模块,用于获取目标文本;

[0012] 确定模块,用于确定通过选择指令所选中的目标歌曲;

[0013] 获取模块还用于获取根据训练完成的曲调控制模型、所述目标文本和所述目标歌曲的曲调信息所合成的自制歌曲;所述自制歌曲以所述目标文本为歌词;

[0014] 播放模块,用于播放所述自制歌曲。

[0015] 一种计算机可读存储介质,所述计算机可读存储介质上存储有计算机程序,所述计算机程序被处理器执行时,使得所述处理器执行以下步骤:

[0016] 获取目标文本;

[0017] 确定通过选择指令所选中的目标歌曲;

[0018] 获取根据训练完成的曲调控制模型、所述目标文本和所述目标歌曲的曲调信息所合成的自制歌曲;所述自制歌曲以所述目标文本为歌词;

[0019] 播放所述自制歌曲。

[0020] 一种计算机设备,包括存储器和处理器,所述存储器中储存有计算机程序,所述计算机程序被所述处理器执行时,使得所述处理器执行以下步骤:

[0021] 获取目标文本；

[0022] 确定通过选择指令所选中的目标歌曲；

[0023] 获取根据训练完成的曲调控制模型、所述目标文本和所述目标歌曲的曲调信息所合成的自制歌曲；所述自制歌曲以所述目标文本为歌词；

[0024] 播放所述自制歌曲。

[0025] 上述音频合成方法、装置、存储介质和计算机设备，在获取到目标文本，并确定通过选择指令所选中的目标歌曲后，即可获取到通过曲调控制模型将目标文本作为歌词，并结合目标歌曲的曲调信息合成自制歌曲。这样用户即可自主选择歌词和曲调，自定义符合自己期望的歌曲，以播放符合自己期望的歌曲，极大地提高了音频播放的效果。

附图说明

[0026] 图1为一个实施例中音频合成方法的应用环境图；

[0027] 图2为一个实施例中音频合成方法的流程示意图；

[0028] 图3为一个实施例中获取目标文本的界面示意图；

[0029] 图4为一个实施例中合成自制歌曲的原理框图；

[0030] 图5为一个实施例中录制视频的界面示意图；

[0031] 图6为一个实施例中提供互动入口的界面示意图；

[0032] 图7为一个实施例中模拟视频通话的界面示意图；

[0033] 图8为一个实施例中音频合成装置的模块结构图；

[0034] 图9为另一个实施例中音频合成装置的模块结构图；

[0035] 图10为一个实施例中计算机设备的内部结构图。

具体实施方式

[0036] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0037] 图1为一个实施例中音频合成方法的应用环境图。参照图1，该音频合成方法应用于音频合成系统。该音频合成系统包括终端110和服务器120。其中，终端110和服务器120通过网络连接。终端110具体可以是台式终端或移动终端，移动终端具体可以手机、平板电脑、笔记本电脑等中的至少一种。服务器120具体可以是独立的服务器，也可以是多个独立的服务器组成的服务器集群。

[0038] 终端110可用于执行该音频合成方法。具体地，终端110可获取目标文本，并确定通过选择指令所选中的目标歌曲；然后在本地根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲；其中，该自制歌曲以获取的目标文本为歌词。终端110便可播放该自制歌曲，还可将该自制歌曲配置为背景音频，基于该背景音频录制视频。这里的曲调控制模型可以是由服务器120训练完成后下发至终端110的，也可以是终端110在本地训练得到的。终端110也可将目标文本和目标歌曲的歌曲标识发送至服务器120，由服务器120来根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲后反馈至终端110。

[0039] 图2为一个实施例中音频合成方法的流程示意图。本实施例主要以该音频合成方法应用于图1中的终端110来举例说明。参照图2,该方法具体包括如下步骤:

[0040] S202,获取目标文本。

[0041] 其中,目标文本是待作为歌词合成新歌曲的文本。具体地,目标文本可以是用户通过终端提供的文本输入框输入的文本,也可以是用户通过选择指令从终端提供的文本模板列表中选中的文本,还可以是其他计算机设备传递的文本。当然,目标文本也可以是其他来源的文本,本发明的实施例中不对目标文本的来源和获取方式进行限定。

[0042] 举例说明,图3示出了一个实施例中获取目标文本的界面示意图。参考图3,该界面包括文本输入框310和文本模板列表320。用户可在文本输入框310中输入文本311,终端从而获取用户在文本输入框310中输入的文本311作为目标文本。用户也可在文本模板列表320中进行文本选择,终端从而获取用户从文本模板列表320中选中的文本321作为目标文本。

[0043] S204,确定通过选择指令所选中的目标歌曲。

[0044] 其中,目标歌曲是待合成的新歌曲的曲调所选自的歌曲。比如,用户想要以“我爱北京天安门”这首歌的曲调作为新歌曲的曲调来合成歌曲,那么“我爱北京天安门”即为目标歌曲。再比如,用户想要以“两只老虎”这首歌的曲调作为新歌曲的曲调来合成歌曲,那么“两只老虎”即为目标歌曲。

[0045] 在一个实施例中,终端可提供候选歌曲,并接收用户对候选歌曲的选择指令,从而获取选择指令所选中的候选歌曲作为目标歌曲。

[0046] 具体地,终端可在界面上展示候选歌曲,继而检测对于展示的候选歌曲的预定义的触发操作,在检测到该触发操作时触发相应的选择指令。触发操作可以是对界面中展示的候选歌曲的触控操作或者光标点击操作等。触发操作也可以是对预定义的物理按钮的点击,或者在逐一显示候选歌曲时触发的晃动操作。

[0047] 在另外的实施例中,终端也可从候选歌曲中随机选取一首歌曲作为目标歌曲。终端也可从候选歌曲中挑选出与目标文本匹配的歌曲作为目标歌曲。

[0048] 举例说明。继续参考图3,该界面还包括候选歌曲列表330。用户可在候选歌曲列表330中进行歌曲选择,终端从而获取用户从候选歌曲列表330中选中的歌曲331作为目标歌曲。

[0049] S206,获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲;自制歌曲以目标文本为歌词。

[0050] 其中,曲调控制模型是经过训练后具有曲调控制能力的机器学习模型。机器学习英文全称为Machine Learning,简称ML。机器学习模型可通过样本学习具备曲调控制能力。机器学习模型可采用神经网络模型、支持向量机或者逻辑回归模型等。神经网络模型比如卷积神经网络、反向传播神经网络、反馈神经网络、径向基神经网络或者自组织神经网络等。本发明的实施例中对机器学习模型的类型不作限定,能实现曲调控制功能的机器学习模型均可。

[0051] 曲调信息是可反映出歌曲的音高随时间变化的趋势的数据。曲调有两个基本要素,即旋律(或称音高)和节奏。终端可通过一系列音高和相应的时间信息的对应关系来表示曲调信息。这里的时间信息用于反映节奏,比如音高的开始时间,或者音高的持续时长,

或者不同音高之间的停顿时长等。一个音高代表一个音符。一个音符的声学参数包括基频和强度等。

[0052] 可以理解,歌曲是由歌词和曲调相结合的一种艺术表现形式。目标歌曲是歌曲演唱者按照歌曲创作者创作的歌词和曲调演唱得到的音频数据。本实施例中,通过保持目标歌曲的曲调不变,更换歌词以合成曲调不变的自制歌曲。当然,本发明实施例中也可根据更换的歌词调整目标歌曲的曲调来合成自制歌曲。

[0053] 其中,自制歌曲是自定义歌词的歌曲。自制歌曲的曲调沿用已有歌曲的曲调。当然,自制歌曲也可以自定义曲调。终端根据原有歌词和自定义曲调合成自制歌曲,或者根据自定义歌词和自定义曲调合成自制歌曲。根据自定义歌词和自定义曲调合成的自制歌曲即为原创歌曲。

[0054] 在一个实施例中,目标歌曲的曲调信息可以是从目标歌曲的歌曲文件中提取的数据。比如,从目标歌曲的歌曲文件中提取MIDI(Musical Instrument Digital Interface,乐器数字接口)文件作为目标歌曲的曲调信息。可以理解,由于MIDI文件本身包括音高和相应的时间信息,而曲调信息可以由一系列音高和相应的时间信息的对应关系来表示,因此MIDI文件本身所包括的音高和相应的时间信息就可以构成曲调信息。

[0055] 在一个实施例中,目标歌曲的曲调信息是从歌曲演唱者演唱的音频数据中提取的。歌曲演唱者演唱的音频数据可以是携带背景伴奏的音频数据,也可以是清唱的音频数据,清唱的音频数据是记录了歌曲演唱者的演唱声音但没有背景音乐的数据,终端可录制与目标歌曲对应的歌曲演唱者演唱的音频数据,再从中提取曲调信息。其中,曲调信息可以是通过构建好的模型来提取。当然,目标歌曲的曲调信息也可以通过其他方式得到,在本发明的实施例中不作限定。

[0056] 具体地,终端在确定通过选择指令所选中的目标歌曲后,可查询该选中的目标歌曲的曲调信息,继而将目标文本和查询到的曲调信息共同输入曲调控制模型,以通过曲调控制模型输出以目标文本为歌词、且以目标歌曲的曲调为曲调的自制歌曲。

[0057] 在一个实施例中,S206包括:将目标文本和目标歌曲的歌曲标识发送至服务器;目标文本和歌曲标识,用于指示服务器在查找到与歌曲标识对应的曲调信息后,根据训练完成的曲调控制模型、目标文本和曲调信息合成自制歌曲;接收服务器所反馈的自制歌曲。

[0058] 其中,歌曲标识用于唯一标识一个歌曲。具体地,终端在确定通过选择指令所选中的目标歌曲后,可将该目标歌曲的歌曲标识和获取的目标文本一起发送至服务器。这样,服务器在接收到歌曲标识后即查找与该歌曲标识对应的曲调信息,将该曲调信息和目标文本输入曲调控制模型,通过曲调控制模型输出以目标文本为歌词、且曲调与目标歌曲一致的自制歌曲。服务器便可将该自制歌曲反馈至终端。

[0059] 其中,自制歌曲可以是自制清唱歌曲,也可以是携带背景伴奏的自制歌曲。当自制歌曲为自制清唱歌曲时,服务器直接将自制清唱歌曲反馈至终端,也可将自制清唱歌曲与背景伴奏混音后得到的自制歌曲反馈至终端,也可将两者均反馈至终端,供终端展示给用户选择。

[0060] 在本实施例中,在服务器侧实现音频合成的步骤,利用服务器强大和高速的计算能力,提高了音频合成的效率。

[0061] S208,播放自制歌曲。

[0062] 具体地,终端在合成自制歌曲或者接收到服务器反馈的自制歌曲后,可播放该自制歌曲,用户即可通过收听播放的自制歌曲,判断是否符合自己的预期,以在自制歌曲符合自己预期时进行后续的操作。

[0063] 上述音频合成方法,在获取到目标文本,并确定通过选择指令所选中的目标歌曲后,即可获取到通过曲调控制模型将目标文本作为歌词,并结合目标歌曲的曲调信息合成自制歌曲。这样用户即可自主选择歌词和曲调,自定义符合自己期望的歌曲,以播放符合自己期望的歌曲,极大地提高了音频播放的效果。

[0064] 在一个实施例中,该音频合成方法还包括:将自制歌曲配置为背景音频,基于该背景音频录制视频。

[0065] 具体地,终端可在合成自制歌曲,或者接收到服务器反馈的自制歌曲后,可自动将该自制歌曲配置为背景音频,并调用本地的图像采集设备以准备录制视频。终端也可根据用户触发的指令将合成的自制歌曲配置为背景音频,再调用本地的图像采集设备以准备录制视频。这样,终端可在合成的自制歌曲作为背景音频下录制视频。

[0066] 在一个具体的实施例中,终端上安装有视频录制应用。终端可根据用户指令运行该视频录制应用,并继续根据用户指令进入背景音频配置界面(例如,可以是如图3所示的界面)。终端可再获取用户输入的目标文本以及用户选取的目标歌曲,进而确定目标歌曲的曲调信息,将目标文本作为歌词,并结合确定的曲调信息合成自制歌曲,终端可再配置合成的制歌曲为背景音频,进入视频录制页面,并根据用户指令播放配置的背景音频并采集图像录制视频。

[0067] 在本实施例中,用户自主选择歌词和曲调,自定义符合自己期望的歌曲,进而将自定义的歌曲作为背景音频来录制视频,极大地丰富了背景音频的来源,且提高了视频录制的效率。

[0068] 在一个实施例中,S206包括:查询与目标歌曲匹配的曲调信息;将目标文本和曲调信息输入曲调控制模型,通过曲调控制模型的隐层根据曲调信息确定目标文本中各字符各自相应的曲调特征;通过曲调控制模型的输出层,输出将目标文本中各字符按照相应的曲调特征进行语音合成得到的自制歌曲。

[0069] 具体地,终端确定通过选择指令所选中的目标歌曲后,查询事先得到的该目标歌曲的曲调信息,再将目标文本和曲调信息共同输入训练完成的曲调控制模型,通过曲调控制模型的隐层对目标文本中的各字符进行曲调匹配,确定目标文本中的各字符各自相应的曲调特征,再通过曲调控制模型的输出层,输出将目标文本中各字符按照相应的曲调特征进行语音合成得到的自制歌曲。

[0070] 在一个实施例中,该音频合成方法还包括训练曲调控制模型的步骤。训练曲调控制模型包括:收集候选歌曲所对应的候选歌曲音频;根据收集的候选歌曲音频,确定各候选歌曲所对应的候选歌曲曲调;获取文本样本;根据文本样本和候选歌曲曲调训练得到所述曲调控制模型。

[0071] 具体地,终端可对各候选歌曲分别录制歌曲演唱者演唱的音频数据,对音频数据进行曲调特征标注,得到相应候选歌曲的曲调信息;或者,获取各候选歌曲的MIDI文件作为相应候选歌曲的曲调信息。终端可再随机选取字符串作为文本样本,将文本样本和候选歌曲的曲调信息共同作为曲调控制模型的训练样本,通过曲调控制模型中的各层对文本样本

和候选歌曲进行处理,以输出以文本样本为歌词、且曲调符合输入的曲调信息的歌曲样本。终端可再根据输出的歌曲样本与期望结果的差异来调整曲调控制模型的模型参数,及作为输入的候选歌曲的曲调信息。终端在调整参数后,可继续训练,直至曲调控制模型训练完成。终端在训练得到曲调控制模型后,可将该曲调控制模型保存在本地,以在需要使用曲调控制模型时获取并使用。

[0072] 在一个实施例中,曲调控制模型输出的歌曲样本与期望结果的差异可通过人工判断。曲调控制模型是对各候选歌曲通用的模型。

[0073] 在一个实施例中,模型训练的过程也可以在服务器上进行,服务器在训练完成曲调控制模型后,将曲调控制模型下发至终端使用。

[0074] 上述实施例中,通过机器学习模型强大的学习能力学习曲调控制的能力,通过训练完成的曲调控制模型进行语音合成,相较于传统的语音合成方式,提高了语音合成效率,而且提高了语音合成的准确性。

[0075] 在一个实施例中,该音频合成方法还包括:确定目标说话对象。S206包括:查询与目标说话对象对应的曲调控制模型;根据查询到的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲;自制歌曲的音色符合目标说话对象。

[0076] 其中,目标说话对象是目标音色所属的对象。目标音色是用户意图合成的自制歌曲或者自制语音的声音特征。目标说话对象可以是虚拟对象,也可以是真实对象。虚拟对象比如游戏中的虚拟角色或者动漫中的虚拟角色等。真实对象比如现实场景中的歌唱家或者演员等。

[0077] 可以理解,不同的发声体由于材料或者结构不同,导致发出的声音的音色也不同。例如钢琴、小提琴和人的音色不一样,发出的声音也就不一样。每个人的音色也会不一样。音色是声音的特点。不同的音色,即使在同一音高和同一声音强度下,也有所区别。

[0078] 需要说明的是,目标说话对象对应的曲调控制模型,是根据由目标说话人演唱的音频数据训练得到的曲调控制模型,这样训练得到的曲调控制模型所输出的自制歌曲的音色符合目标说话对象。而前述实施例中所提到的曲调控制模型不限制训练使用的音频数据的演唱者,这样即不对音色作出限定。

[0079] 具体地,终端可确定通过选择指令继续选中的目标说话对象,终端也可获取目标歌曲的演唱者作为目标说话对象。进一步地,终端再将目标文本和目标歌曲的曲调信息输入与目标说话对象对应的曲调控制模型合成自制歌曲。

[0080] 可以理解,与目标说话对象对应的曲调控制模型,即为与目标说话对象的音色信息匹配的曲调控制模型,终端将目标文本和曲调信息共同输入训练完成的、且与目标说话对象的音色信息匹配曲调控制模型,通过曲调控制模型的隐层对目标文本中的各字符进行曲调匹配,确定各字符各自相应的曲调特征,通过曲调控制模型的输出层,输出将目标文本中各字符按照相应的曲调特征进行语音合成得到的、且符合目标说话对象音色的自制歌曲。

[0081] 当然,终端在通过模型实现该过程时,可以通过一个模型实现端到端(目标文本到自制歌曲)的转换,也可以通过多个模型实现,各模型分工合作,各自实现一个功能。比如,先通过音色控制模型将目标文本转化为音色符合目标说话对象的自制语音,再通过曲调控制模型将自制语音转化为曲调符合目标歌曲的自制歌曲。这里的音色控制模型可参考后文

描述,这里的曲调控制模型训练输入数据不再是目标文本和曲调信息,而是音色控制模型输出的自制语音和曲调信息。

[0082] 可以理解,本实施例可以合成通过用户选择的声音唱出用户自定义歌词的歌曲。比如某个明星的声音唱出用户自制歌曲,以完成歌曲创作。

[0083] 在一个实施例中,终端还可对获取的目标文本进行安全检测,在目标文本符合规定时,继续后续的操作;在目标文本不符合规定时,则反馈错误提示。其中,目标文本的安全检测步骤也可由服务器进行。

[0084] 举例说明,图4示出了一个实施例中合成自制歌曲的原理框图。参考图4,服务器(或终端)可事先收集声音素材(候选说话对象的音频数据),对收集的声音素材进行标注(音素、基频、声音强度以及时间信息等),并根据收集的声音素材和标注数据训练模型。这里的模型可以是音色控制模型,或者通用的曲调控制模型,或者与各候选说话对象对应的曲调控制模型等;根据从声音素材选取的作为模型输入的数据和标注数据的不同,可训练出不同的模型。服务器(或终端)可保存训练得到的模型以备使用,服务器也可将模型下发到终端使用。

[0085] 终端在获取目标文本后,可将该目标文本上传至服务器,终端还可将用户选中的目标歌曲或者目标说话对象的标识上传至服务器。由服务器对目标文本进行安全检查。在安全检查通过后,在用户选取的目标说话对象时,通过训练得到的音色控制模型根据目标文本的音素序列得到目标文本对应的自制语音;或者在用户选取的目标歌曲时,通过训练得到的曲调控制模型根据目标歌曲的曲调信息得到以目标文本为歌词的自制歌曲,或者在用户选取目标歌曲和目标说话对象时,通过训练得到的曲调控制模型根据目标文本的音素序列和目标歌曲的曲调信息得到以目标文本为歌词的自制歌曲。服务器在通过模型得到自制歌曲或者自制语音后下发至终端,或者将自制歌曲或者自制语音与伴奏混音后再下发至终端。终端也可以在本地完成上述过程。

[0086] 上述实施例中,用户可自主选择歌词、曲调和音色,自定义符合自己期望的歌曲,进而将自定义的歌曲作为背景音频来录制视频,极大地丰富了背景音频的来源,且提高了视频录制的效率。

[0087] 在一个实施例中,该音频合成方法还包括:接收选择指令;当选择指令选中目标歌曲时,则执行获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲的步骤;当选择指令选中目标说话对象时,则获取根据训练完成的音色控制模型和目标文本所合成的自制音频;音色控制模型与目标说话对象匹配。

[0088] 其中,选择指令是根据用户操作触发的计算机可读指令。选择指令用于对目标进行选取。目标可以是目标歌曲或者目标说话对象。音色控制模型是经过训练后具有音色控制能力的机器学习模型。

[0089] 具体地,终端可在界面上提供候选歌曲列表或者候选说话对象列表,继而检测对于展示的候选歌曲或者候选说话对象的预定义的触发操作,在检测到该触发操作时触发相应的选择指令。终端在确定通过选择指令指向目标歌曲后,执行获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲的步骤。

[0090] 终端在确定通过选择指令指向目标说话对象后,则获取根据训练完成的音色控制模型和目标文本所合成的自制音频;其中,音色控制模型与目标说话对象匹配。可以理解,

本实施例可以实现,用户选择通过某个虚拟角色或者真实人物的声音说出自己想说的话。

[0091] 在一个实施例中,获取根据训练完成的音色控制模型和目标文本所合成的自制音频,包括:查询与目标说话对象匹配的音色控制模型;确定目标文本所对应的音素序列;通过音色控制模型根据音素序列合成自制语音;根据自制语音和背景伴奏合成自制音频。

[0092] 其中,与目标说话对象匹配的音色控制模型,是根据目标说话对象的音频数据训练得到的音色控制模型,通过该音色控制模型输出的语音的音色与目标说话对象的音色一致。

[0093] 音素是根据语音的自然属性划分出的最小语音单元。以汉语音节举例说明:ā(啊)只有一个音素,ài(爱)有两个音素,dāi(呆)则有三个音素等。终端中可存储有字符与音素的对应关系。

[0094] 具体地,终端在获取目标文本后,可根据字符与音素的对应关系,确定目标文本所对应的音素序列终端可,再将目标文本所对应的音素序列输入训练完成的音色控制模型,通过音色控制模型的隐层确定各音素各自对应的声学参数,继而通过音色控制模型的隐层根据各音素各自对应的声学参数,确定各音素各自对应的语音波形单元,从而通过音色控制模型的输出层根据各音素各自对应的语音波形单元,输出目标文本所对应的语音波形,得到自制语音。

[0095] 其中,音色控制模型输出的自制语音是没有背景伴奏的目标说话对象的说话声,终端可将自制语音和背景伴奏混音,得到自制音频。

[0096] 在本实施例中,通过机器学习模型强大的学习能力学习音色控制的能力,通过训练完成的音色控制模型进行语音合成,相较于传统的语音合成方式,提高了语音合成效率,而且提高了语音合成的准确性。

[0097] 在一个实施例中,该音频合成方法还包括训练与各候选说话对象匹配的音色控制模型的步骤;目标说话对象选择候选说话对象;训练与各候选说话对象匹配的音色控制模型,包括:收集各候选说话对象相应的音频素材;根据各音频素材确定相应候选说话对象所相应的音素素材序列;采用各候选说话对象所相应的音素素材序列,训练得到各候选说话对象匹配的音色控制模型。

[0098] 具体地,终端可收集各候选说话对象相应的音频素材。对于每个候选说话对象,对该候选说话对象相应的音频素材进行音频切分,得到音频素材对应的音素序列样本,将音素序列样本作为输入数据训练音色控制模型,以输出符合该候选说话对象的音色的音频。终端可再根据输出的音频与输入的音素序列样本所来自的音频素材的差异来调整音色控制模型的模型参数。终端在调整参数后,可继续训练,直至音色控制模型训练完成。这样,终端即对每个候选说话对象分别训练了相匹配的音色控制模型。

[0099] 终端在训练得到音色控制模型后,可将该音色控制模型保存在本地,以在需要使用音色控制模型时获取并使用。音色控制模型训练的过程也可在服务器上进行,服务器在训练完成音色控制模型后,将音色控制模型下发至终端使用。

[0100] 在一个实施例中,终端可对自制语音进行音节划分,通常一个音节表示一个字符。这样,终端可调整各音节的时间信息,如开始时间、持续时长及不同音节之间的停顿时长等,使得自制语音的节奏符合曲调信息;终端可再调整各音节的音高,如基频和声音强度等,使得自制语音的旋律符合曲调信息;从而得到以目标文本为歌词、曲调与目标歌曲一

致、且音色与目标说话对象一致的自制歌曲。

[0101] 上述本实施例中,用户可以自主选择歌词和曲调,自定义符合自己期望的歌曲,进而将自定义的歌曲作为背景音频来录制视频;或者自主选择目标音色,以得符合自己选中的音色的语音。

[0102] 在一个实施例中,接收选择指令,包括:接收对应于虚拟对象附加元素的选择指令;确定选择指令选中的虚拟对象附加元素所对应的目标说话对象。该音频合成方法还包括:将自制音频配置为背景音频;将虚拟对象附加元素叠加至采集的图像得到视频帧;基于背景音频和叠加得到的视频帧生成录制的视频。

[0103] 其中,虚拟对象附加元素是可视内容为虚拟对象形象的数据。虚拟对象附加元素可以用于额外添加在图像中来修饰图像内容。比如虚拟对象挂件等。虚拟对象附加元素可以是动态数据,比如动态图片;也可以是静态数据,比如静态图片。

[0104] 具体地,终端可在界面上提供虚拟对象附加元素列表,继而检测对于展示的虚拟对象附加元素的预定义的触发操作,在检测到该触发操作时触发相应的选择指令,并选中选择指令指向的虚拟对象附加元素。终端继而将选中虚拟对象附加元素所对应的虚拟对象作为目标说话对象。进一步地,终端可获取目标文本,根据该目标说话对象匹配的音色控制模型,将目标文本转换为自制音频,再通过图像采集装置采集图像,将虚拟对象附加元素叠加至采集的图像得到视频帧。终端也可渲染得到的视频帧形成预览画面,基于自制音频录制视频。

[0105] 举例说明,图5示出了一个实施例中录制视频的界面示意图。参考图5左图,该界面包括虚拟对象附加元素列表510和文本输入框入口520。用户可在虚拟对象附加元素列表510中进行选择,终端从而获取用户从虚拟对象附加元素列表510中选中的虚拟对象附加元素511,将虚拟对象附加元素所对应的虚拟对象作为目标说话对象。用户可通过文本输入框入口520开启文本输入框530,如图5中图所示,在文本输入框中输入目标文本,或者如图5右图所示,在候选文本模板中选择目标文本。

[0106] 在具体的场景下,终端可在视频录制应用的拍摄界面上,提供游戏的英雄挂件,并可以让用户自定义一段文本,然后就可以用这个英雄的声音把用户自定义的文本念出来,达到用户直接指挥游戏英雄说出想说的话的效果。在合成自制语音后,在以自制语音为背景音频录制视频时,用户可以通过对口型表演,从装扮、声音和口型等多个方位,达到模仿的效果。

[0107] 在一个实施例中,接收选择指令,包括:接收用于进行模拟视频通话的选择指令;确定选择指令选中的目标说话对象所对应的图片。该音频合成方法还包括:将自制音频配置为背景音频;根据图片和采集的图像生成通话视频帧;基于背景音频和生成的通话视频帧生成录制的视频。

[0108] 具体地,终端可在视频播放界面提供互动入口。终端可根据用户指令通过该互动入口进行互动界面。终端可在互动界面提供模拟视频通话入口并,在接收到作用于模拟视频通话入口的选择指令后,确定选择指令选中的目标说话对象为当前播放的视频中的对象,或者发布该视频的用户;再确定目标说话对象所对应的图片,根据确定的图片和采集的图像生成通话视频帧,基于背景音频和生成的通话视频帧生成与目标说话对象的进行视频通话的视频。

[0109] 终端还可在互动界面提供文本输入框,获取输入的目标文本,根据确定的目标说话对象的音色信息,将目标文本转换为自制语音,以该自制语音和生成的通话视频帧生成录制的视频。这样,即可达到用户直接可控制通话视频对象说出用户想要视频通话对象对自己说的话的效果。

[0110] 举例说明,图6示出了一个实施例中提供互动入口的界面示意图。参考图6左图,该界面包括当前播放视频的视频帧610和互动入口620。终端可根据用户指令通过该互动入口620进行如图6右图所示的互动界面。该互动界面中包括文本输入框631和模拟视频通话入口632。用户可通过模拟视频通话入口632模拟与视频帧610中的对象进行视频通话。用户可通过文本输入框631输入文本,终端则获取用户输入的文本为目标文本,来合成自制语音,将该自制语音在模拟视频通话时播放,作为模拟视频通话内容。

[0111] 图7示出了一个实施例中模拟视频通话的界面示意图。参考图7左图,为视频通话邀请界面。视频通话邀请界面中包括目标说话对象所对应的图片711。用户可通过视频通话接听入口进行模拟视频通话。终端也可在渲染视频通话邀请界面预设时长后进入视频通话界面。再参考图7右图,为视频通话界面,该界面包括目标说话对象(视频对象)所对应的图片721,和采集的图像722,模拟与目标说话对象进行视频通话的场景。

[0112] 在具体的场景下,终端可在视频录制应用提供“明星说出你的名字”的玩法。用户在看视频时,会出现引导输入名字的弹窗,根据指引可以输入2-4个字的文本,输入的文本会先给到服务器做安全鉴定,文本符合规定,则将用户输入的文本和明星语料库做语音合成处理,合成好的自制语音返回到终端,终端响应跳转行为,跳到拍摄页,加载明星打电话的素材(视频通话邀请界面)。素材加载成功后,就可以看到明星来电的页面,明星会在视频中念出刚输入的文本(即播放自制语音),同时用户会以小窗的形式,出现在拍摄界面的右上方,达到跟明星同屏的效果,按下拍摄按钮,就可以获得一条独一无二的明星同屏视频。

[0113] 应该理解的是,虽然上述各实施例的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,这些步骤可以以其它的顺序执行。而且,上述各实施例中的至少一部分步骤可以包括多个子步骤或者多个阶段,这些子步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,这些子步骤或者阶段的执行顺序也不必然是依次进行,而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0114] 如图8所示,在一个实施例中,提供了一种音频合成装置800。参照图8,该音频合成装置800包括:获取模块801、确定模块802和播放模块803。

[0115] 获取模块801,用于获取目标文本。

[0116] 确定模块802,用于确定通过选择指令所选中的目标歌曲。

[0117] 获取模块801还用于获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲;自制歌曲以目标文本为歌词。

[0118] 播放模块803,用于播放自制歌曲。

[0119] 在一个实施例中,获取模块801还用于将目标文本和目标歌曲的歌曲标识发送至服务器;目标文本和歌曲标识,用于指示服务器在查找到与歌曲标识对应的曲调信息后,根据训练完成的曲调控制模型、目标文本和曲调信息合成自制歌曲;接收服务器所反馈的自

制歌曲。

[0120] 在一个实施例中,获取模块801还用于查询与目标歌曲匹配的曲调信息;将目标文本和曲调信息输入曲调控制模型,通过曲调控制模型的隐层根据曲调信息确定目标文本中各字符各自相应的曲调特征;通过曲调控制模型的输出层,输出将目标文本中各字符按照相应的曲调特征进行语音合成得到的自制歌曲。

[0121] 在一个实施例中,音频合成装置800还包括训练模块804用于训练曲调控制模型;具体用于收集候选歌曲所对应的候选歌曲音频;根据收集的候选歌曲音频,确定各候选歌曲所对应的候选歌曲曲调;获取文本样本;根据文本样本和候选歌曲曲调训练得到曲调控制模型。其中,目标歌曲选自候选歌曲。

[0122] 在一个实施例中,确定模块802还用于确定目标说话对象;获取模块801还用于查询与目标说话对象对应的曲调控制模型;根据查询到的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲;自制歌曲的音色符合目标说话对象。

[0123] 如图9所示,在一个实施例中,音频合成装置800还包括训练模块804和录制模块805。

[0124] 录制模块805用于将自制歌曲配置为背景音频;基于背景音频录制视频。

[0125] 在一个实施例中,获取模块801还用于接收选择指令;当选择指令选中目标歌曲时,获取模块801还用于获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲;当选择指令选中目标说话对象时,获取模块801还用于获取根据训练完成的音色控制模型和目标文本所合成的自制音频;音色控制模型与目标说话对象匹配。

[0126] 在一个实施例中,训练模块804还用于训练与各候选说话对象匹配的音色控制模型;具体用于收集各候选说话对象相应的音频素材;根据各音频素材确定相应候选说话对象所相应的音素素材序列;采用各候选说话对象所相应的音素素材序列,训练得到各候选说话对象匹配的音色控制模型。其中,目标说话对象选自候选说话对象。

[0127] 在一个实施例中,获取模块801还用于查询与目标说话对象匹配的音色控制模型;确定目标文本所对应的音素序列;通过音色控制模型根据音素序列合成自制语音;根据自制语音和背景伴奏合成自制音频。

[0128] 在一个实施例中,获取模块801还用于接收对应于虚拟对象附加元素的选择指令;确定选择指令选中的虚拟对象附加元素所对应的目标说话对象。录制模块805还用于将自制音频配置为背景音频;将虚拟对象附加元素叠加至采集的图像得到视频帧;基于背景音频和叠加得到的视频帧生成录制的视频。

[0129] 在一个实施例中,获取模块801还用于接收用于进行模拟视频通话的选择指令;确定选择指令选中的目标说话对象所对应的图片。录制模块805还用于将自制音频配置为背景音频;根据图片和采集的图像生成通话视频帧;基于背景音频和生成的通话视频帧生成录制的视频。

[0130] 图10示出了一个实施例中计算机设备的内部结构图。该计算机设备具体可以是图1中的终端110。如图10所示,该计算机设备包括通过系统总线连接的处理器、存储器、网络接口、输入设备和显示屏。其中,存储器包括非易失性存储介质和内存储器。该计算机设备的非易失性存储介质存储有操作系统,还可存储有计算机程序,该计算机程序被处理器执行时,可使得处理器实现音频合成方法。该内存储器中也可储存有计算机程序,该计算机程

序被处理器执行时,可使得处理器执行音频合成方法。计算机设备的显示屏可以是液晶显示屏或者电子墨水显示屏等,输入装置可以是显示屏上覆盖的触摸层,也可以是计算机设备外壳上设置的按键、轨迹球或触控板,也可以是外接的键盘、触控板或鼠标等。本领域技术人员可以理解,图10中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备的限定,具体的计算机设备可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0131] 在一个实施例中,本申请提供的音频合成装置可以实现为一种计算机程序的形式,计算机程序可在如图10所示的计算机设备上运行,计算机设备的非易失性存储介质可存储组成该音频合成装置的各个程序模块,比如,图8所示的获取模块801、确定模块802、合成模块803、配置模块804和录制模块805等。各个程序模块组成的计算机程序使得处理器执行本说明书中描述的本申请各个实施例的音频合成方法中的步骤。

[0132] 例如,图10所示的计算机设备可以通过如图8所示的音频合成装置800中的获取模块801获取目标文本。通过确定模块802确定通过选择指令所选中的目标歌曲。通过获取模块801获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲;自制歌曲以目标文本为歌词。通过播放模块803播放自制歌曲。

[0133] 在一个实施例中,提供了一种计算机可读存储介质,该计算机可读存储介质上存储有计算机程序,该计算机程序被处理器执行时,使得处理器执行以下步骤:获取目标文本;确定通过选择指令所选中的目标歌曲;获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲;自制歌曲以目标文本为歌词;播放自制歌曲。

[0134] 在一个实施例中,获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲,包括:将目标文本和目标歌曲的歌曲标识发送至服务器;目标文本和歌曲标识,用于指示服务器在查找到与歌曲标识对应的曲调信息后,根据训练完成的曲调控制模型、目标文本和曲调信息合成自制歌曲;接收服务器所反馈的自制歌曲。

[0135] 在一个实施例中,获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲,包括:查询与目标歌曲匹配的曲调信息;将目标文本和曲调信息输入曲调控制模型,通过曲调控制模型的隐层根据曲调信息确定目标文本中各字符各自相应的曲调特征;通过曲调控制模型的输出层,输出将目标文本中各字符按照相应的曲调特征进行语音合成得到的自制歌曲。

[0136] 在一个实施例中,该计算机程序还使得处理器执行以下步骤:训练曲调控制模型;其中,目标歌曲选自候选歌曲。训练曲调控制模型,包括:收集候选歌曲所对应的候选歌曲音频;根据收集的候选歌曲音频,确定各候选歌曲所对应的候选歌曲曲调;获取文本样本;根据文本样本和候选歌曲曲调训练得到曲调控制模型。

[0137] 在一个实施例中,该计算机程序还使得处理器执行以下步骤:确定目标说话对象。获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲,包括:查询与目标说话对象对应的曲调控制模型;根据查询到的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲;自制歌曲的音色符合目标说话对象。

[0138] 在一个实施例中,该计算机程序还使得处理器执行以下步骤:将自制歌曲配置为背景音频;基于背景音频录制视频。

[0139] 在一个实施例中,该计算机程序还使得处理器执行以下步骤:接收选择指令;当选

择指令选中目标歌曲时，则执行获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲的步骤；当选择指令选中目标说话对象时，则获取根据训练完成的音色控制模型和目标文本所合成的自制音频；音色控制模型与目标说话对象匹配。

[0140] 在一个实施例中，该计算机程序还使得处理器执行以下步骤：训练与各候选说话对象匹配的音色控制模型；其中，目标说话对象选自候选说话对象。训练与各候选说话对象匹配的音色控制模型，包括：收集各候选说话对象相应的音频素材；根据各音频素材确定相应候选说话对象所相应的音素素材序列；采用各候选说话对象所相应的音素素材序列，训练得到各候选说话对象匹配的音色控制模型。

[0141] 在一个实施例中，获取根据训练完成的音色控制模型和目标文本所合成的自制音频，包括：查询与目标说话对象匹配的音色控制模型；确定目标文本所对应的音素序列；通过音色控制模型根据音素序列合成自制语音；根据自制语音和背景伴奏合成自制音频。

[0142] 在一个实施例中，接收选择指令，包括：接收对应于虚拟对象附加元素的选择指令；确定选择指令选中的虚拟对象附加元素所对应的目标说话对象。该计算机程序还使得处理器执行以下步骤：将自制音频配置为背景音频；将虚拟对象附加元素叠加至采集的图像得到视频帧；基于背景音频和叠加得到的视频帧生成录制的视频。

[0143] 在一个实施例中，接收选择指令，包括：接收用于进行模拟视频通话的选择指令；确定选择指令选中的目标说话对象所对应的图片。该计算机程序还使得处理器执行以下步骤：将自制音频配置为背景音频；根据图片和采集的图像生成通话视频帧；基于背景音频和生成的通话视频帧生成录制的视频。

[0144] 在一个实施例中，提供了一种计算机设备，包括存储器和处理器，存储器中储存有计算机程序，计算机程序被处理器执行时，使得处理器执行以下步骤：获取目标文本；确定通过选择指令所选中的目标歌曲；获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲；自制歌曲以目标文本为歌词；播放自制歌曲。

[0145] 在一个实施例中，获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲，包括：将目标文本和目标歌曲的歌曲标识发送至服务器；目标文本和歌曲标识，用于指示服务器在查找到与歌曲标识对应的曲调信息后，根据训练完成的曲调控制模型、目标文本和曲调信息合成自制歌曲；接收服务器所反馈的自制歌曲。

[0146] 在一个实施例中，获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲，包括：查询与目标歌曲匹配的曲调信息；将目标文本和曲调信息输入曲调控制模型，通过曲调控制模型的隐层根据曲调信息确定目标文本中各字符各自相应的曲调特征；通过曲调控制模型的输出层，输出将目标文本中各字符按照相应的曲调特征进行语音合成得到的自制歌曲。

[0147] 在一个实施例中，该计算机程序还使得处理器执行以下步骤：训练曲调控制模型；其中，目标歌曲选自候选歌曲。训练曲调控制模型，包括：收集候选歌曲所对应的候选歌曲音频；根据收集的候选歌曲音频，确定各候选歌曲所对应的候选歌曲曲调；获取文本样本；根据文本样本和候选歌曲曲调训练得到曲调控制模型。

[0148] 在一个实施例中，该计算机程序还使得处理器执行以下步骤：确定目标说话对象。获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲，包括：查询与目标说话对象对应的曲调控制模型；根据查询到的曲调控制模型、目标文本和

目标歌曲的曲调信息所合成的自制歌曲;自制歌曲的音色符合目标说话对象。

[0149] 在一个实施例中,该计算机程序还使得处理器执行以下步骤:将自制歌曲配置为背景音频;基于背景音频录制视频。

[0150] 在一个实施例中,该计算机程序还使得处理器执行以下步骤:接收选择指令;当选择指令选中目标歌曲时,则执行获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲的步骤;当选择指令选中目标说话对象时,则获取根据训练完成的音色控制模型和目标文本所合成的自制音频;音色控制模型与目标说话对象匹配。

[0151] 在一个实施例中,该计算机程序还使得处理器执行以下步骤:训练与各候选说话对象匹配的音色控制模型;其中,目标说话对象选自候选说话对象。训练与各候选说话对象匹配的音色控制模型,包括:收集各候选说话对象相应的音频素材;根据各音频素材确定相应候选说话对象所相应的音素素材序列;采用各候选说话对象所相应的音素素材序列,训练得到各候选说话对象匹配的音色控制模型。

[0152] 在一个实施例中,获取根据训练完成的音色控制模型和目标文本所合成的自制音频,包括:查询与目标说话对象匹配的音色控制模型;确定目标文本所对应的音素序列;通过音色控制模型根据音素序列合成自制语音;根据自制语音和背景伴奏合成自制音频。

[0153] 在一个实施例中,接收选择指令,包括:接收对应于虚拟对象附加元素的选择指令;确定选择指令选中的虚拟对象附加元素所对应的目标说话对象。该计算机程序还使得处理器执行以下步骤:将自制音频配置为背景音频;将虚拟对象附加元素叠加至采集的图像得到视频帧;基于背景音频和叠加得到的视频帧生成录制的视频。

[0154] 在一个实施例中,接收选择指令,包括:接收用于进行模拟视频通话的选择指令;确定选择指令选中的目标说话对象所对应的图片。该计算机程序还使得处理器执行以下步骤:将自制音频配置为背景音频;根据图片和采集的图像生成通话视频帧;基于背景音频和生成的通话视频帧生成录制的视频。

[0155] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的程序可存储于一非易失性计算机可读取存储介质中,该程序在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0156] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0157] 以上实施例仅表达了本发明的几种实施方式,其描述较为具体和详细,但并不能因此而理解为对本发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本发明构思的前提下,还可以做出若干变形和改进,这些都属于本发明的保护范

围。因此,本发明专利的保护范围应以所附权利要求为准。

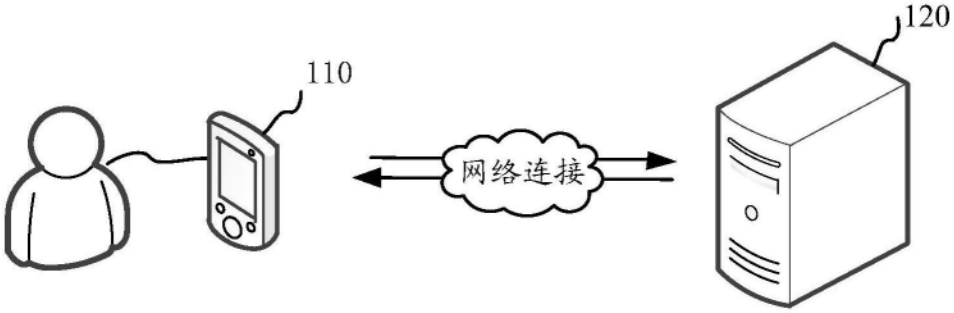

110
120
网络连接

图1

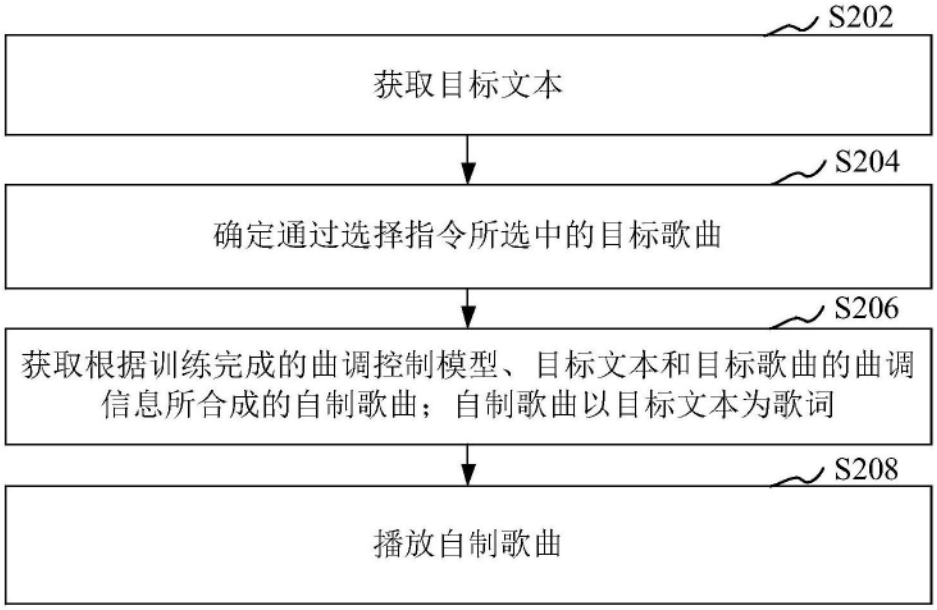

S202
获取目标文本
S204
确定通过选择指令所选中的目标歌曲
S206
获取根据训练完成的曲调控制模型、目标文本和目标歌曲的曲调信息所合成的自制歌曲；自制歌曲以目标文本为歌词
S208
播放自制歌曲

图2

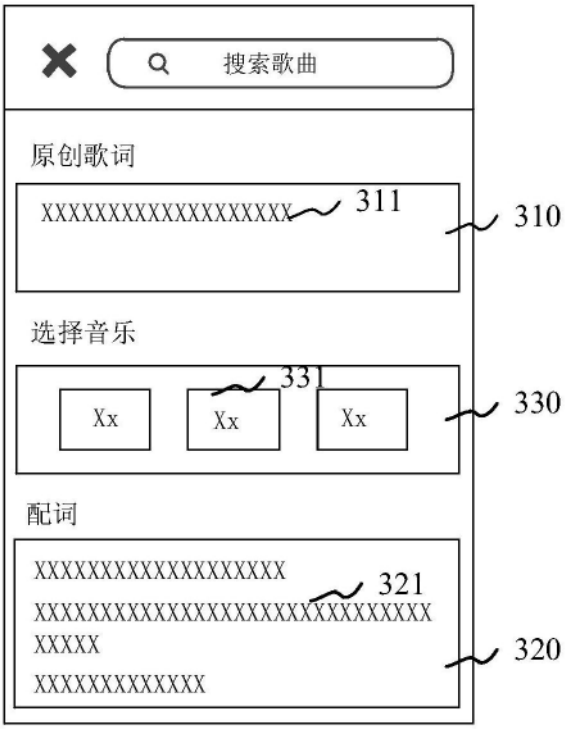

搜索歌曲
原创歌词
XXXXXXXXXXXXXXXXXXX
311
310
选择音乐
331
Xx
Xx
Xx
330
配词
XXXXXXXXXXXXXXXXXXX
321
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXX
XXXXXXXXXXXXXX
320

图3

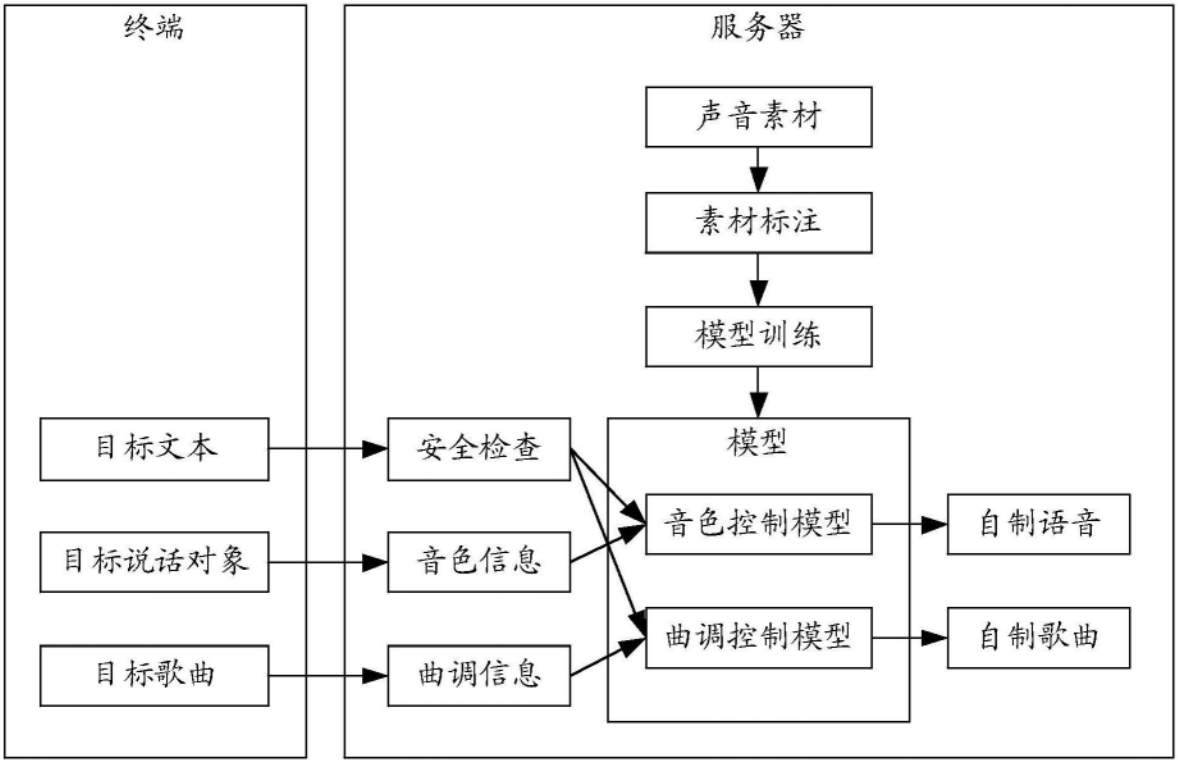

终端
服务器
声音素材
素材标注
模型训练
模型
目标文本
目标说话对象
目标歌曲
安全检查
音色信息
曲调信息
音色控制模型
曲调控制模型
自制语音
自制歌曲

图4

511
520
510
530

图5

关注
推荐
频道
610
620
输入我的名字
631
输入我的名字
2~6个汉字
632
开始打电话

图6

711
XXX
正在给你拨打视频电话
722
721
XXX

图7

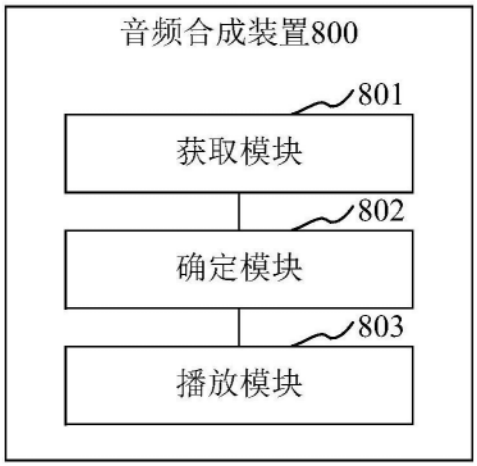

音频合成装置800
801
获取模块
802
确定模块
803
播放模块

图8

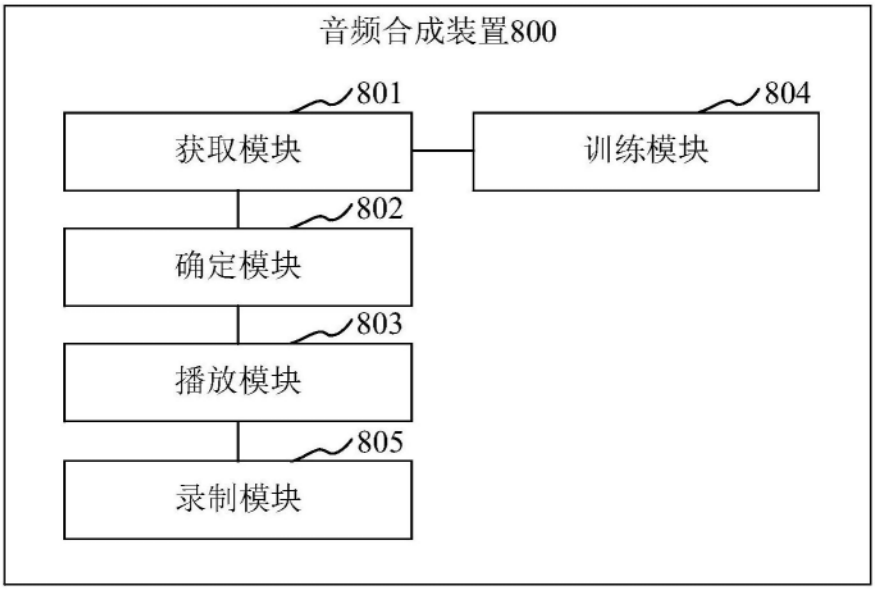

音频合成装置800
801
获取模块
804
训练模块
802
确定模块
803
播放模块
805
录制模块

图9

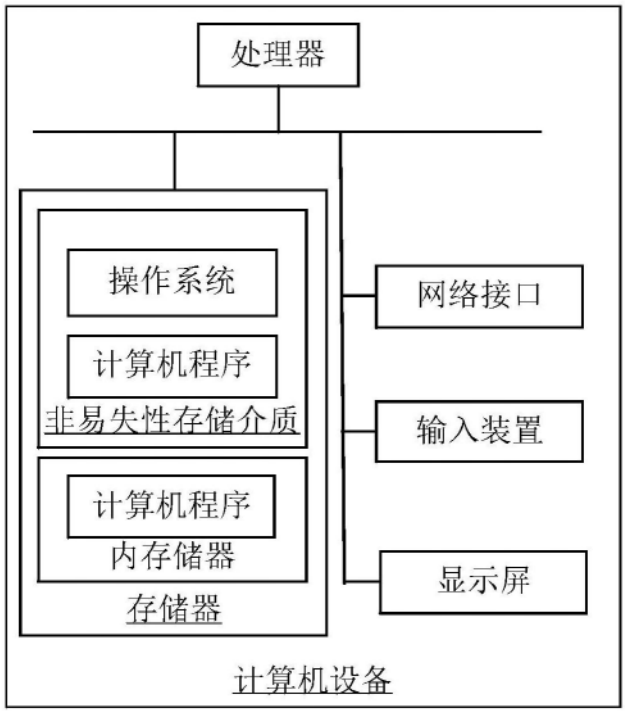

处理器
操作系统
计算机程序
非易失性存储介质
计算机程序
内存储器
存储器
网络接口
输入装置
显示屏
计算机设备

图10