(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116959447 A
(43)申请公布日 2023.10.27

(21)申请号 202211455842.7

(22)申请日 2022.11.21

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

(72)发明人 杨培基

(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
专利代理师 徐耿铭

(51)Int.Cl.
G10L 15/26(2006.01)
G10L 15/06(2013.01)
G10L 25/24(2013.01)
G10L 13/08(2013.01)

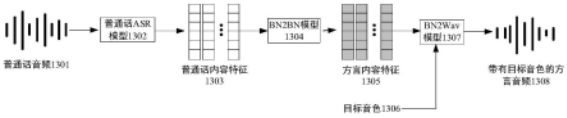
权利要求书3页 说明书15页 附图6页

(54)发明名称

语音转换模型的训练方法、装置、设备及介质

(57)摘要

本申请公开了一种语音转换模型的训练方法、装置、设备及介质。包括:基于第一样本音频训练第一ASR模型,以及基于第二样本音频训练第二ASR模型;基于第一样本音频对应的第一样本文本以及第一样本内容特征,训练第一转换模型,第一转换模型用于将文本转换为第一口音的内容特征;基于第一转换模型、第二样本音频对应的第二样本文本以及第二样本内容特征,构建平行样本数据;基于平行样本数据训练第二转换模型,第二转换模型用于对第一口音和第二口音间进行内容特征转换;基于不同样本音频的样本内容特征训练第三转换模型,第三转换模型用于将内容特征转换为音频;基于训练得到的第一ASR模型、第二转换模型和第三转换模型生成语音转换模型。



CN 116959447 A

1.一种语音转换模型的训练方法,其特征在于,所述方法包括:

基于第一样本音频训练第一ASR模型,以及基于第二样本音频训练第二ASR模型,所述第一样本音频对应第一口音,所述第二样本音频对应第二口音;

基于所述第一样本音频对应的第一样本文本以及第一样本内容特征,训练第一转换模型,所述第一样本内容特征由所述第一ASR模型对所述第一样本音频进行提取得到,所述第一转换模型用于将文本转换为所述第一口音的内容特征;

基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征,构建平行样本数据,所述第二样本内容特征由所述第二ASR模型对所述第二样本音频进行提取得到,所述平行样本数据由不同内容特征构成,不同内容特征对应不同口音,且不同内容特征对应相同文本;

基于所述平行样本数据训练第二转换模型,所述第二转换模型用于对所述第一口音和所述第二口音间进行内容特征转换;

基于不同样本音频的样本内容特征训练第三转换模型,所述第三转换模型用于将内容特征转换为音频;

基于训练得到的所述第一ASR模型、所述第二转换模型和所述第三转换模型生成语音转换模型,所述语音转换模型用于将第一口音的音频转换为第二口音的音频。

2.根据权利要求1所述的方法,其特征在于,所述基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征,构建平行样本数据,包括:

通过所述第一转换模型对所述第二样本文本进行转换,得到第三样本内容特征,所述第三样本内容特征指采用所述第一口音表述第二样本文本所产生音频的内容特征;

基于所述第二样本内容特征和所述第三样本内容特征构建所述平行样本数据。

3.根据权利要求2所述的方法,其特征在于,所述基于所述平行样本数据训练第二转换模型,包括:

将所述第三样本内容特征输入所述第二转换模型,得到第二预测内容特征;

以所述第二样本内容特征为所述第二预测内容特征的监督,训练所述第二转换模型。

4.根据权利要求1所述的方法,其特征在于,所述基于所述第一样本音频对应的第一样本文本以及第一样本内容特征,训练第一转换模型,包括:

将所述第一样本文本输入所述第一转换模型,得到所述第一转换模型输出的第一预测内容特征;

以所述第一样本内容特征为所述第一预测内容特征的监督,训练所述第一转换模型。

5.根据权利要求4所述的方法,其特征在于,所述第一转换模型中包括第一转换子模型、时长预测子模型以及第二转换子模型;

所述将所述第一样本文本输入所述第一转换模型,得到所述第一转换模型输出的第一预测内容特征,包括:

通过所述第一转换子模型对所述第一样本文本进行编码,得到第一文本编码特征;

通过所述时长预测子模型对所述第一文本编码特征进行时长预测,得到预测时长,所述预测时长用于表征所述第一样本文本的发音时长;

基于所述预测时长对所述第一文本编码特征进行特征扩充,得到第二文本编码特征;

通过所述第二转换子模型对所述第二文本编码特征进行转换,得到所述第一预测内容

特征。

6.根据权利要求5所述的方法,其特征在于,所述第一转换子模型和所述第二转换子模型由FFT堆叠而成,所述FFT由多头注意力机制层和卷积层构成。

7.根据权利要求1所述的方法,其特征在于,所述基于不同样本音频的样本内容特征训练第三转换模型,包括:

将所述样本内容特征以及所述样本音频对应的说话者标识输入所述第三转换模型,得到预测音频,所述预测音频与所述样本音频对应相同音频内容,且具有相同音色,其中,不同说话者对应不同说话者标识;

基于所述预测音频以及所述样本音频,训练所述第三转换模型。

8.根据权利要求7所述的方法,其特征在于,所述第三转换模型包括第三转换子模型以及声码器;

所述将所述样本内容特征以及所述样本音频对应的说话者标识输入所述第三转换模型,得到预测音频,包括:

将所述样本内容特征以及所述说话者标识输入所述第三转换子模型,得到预测音频谱特征;

将所述预测音频谱特征输入所述声码器,得到所述预测音频。

9.根据权利要求8所述的方法,其特征在于,所述将所述预测音频谱特征输入所述声码器,得到所述预测音频之前,所述方法还包括:

以所述样本音频的样本音频谱特征为所述预测音频谱特征的监督,训练所述第三转换子模型;

所述将所述预测音频谱特征输入所述声码器,得到所述预测音频,包括:

在所述第三转换子模型训练完成的情况下,将训练完成后所述第三转换子模型输出的所述预测音频谱特征输入所述声码器,得到所述预测音频;

所述基于所述预测音频以及所述样本音频,训练所述第三转换模型,包括:

基于所述预测音频以及所述样本音频,训练所述第三转换模型中的所述声码器。

10.根据权利要求1至9任一所述的方法,其特征在于,所述方法包括:

响应于口音转换指令,通过所述第一ASR模型提取第一口音音频的第一内容特征,第一内容特征对应所述第一口音,所述口音转换指令用于指示将音频由所述第一口音转换为所述第二口音;

通过所述第二转换模型将所述第一内容特征转换为第二内容特征,所述第二内容特征对应所述第二口音;

通过所述第三转换模型对所述第二内容特征进行音频转换,得到第二口音音频。

11.根据权利要求10所述的方法,其特征在于,所述口音转换指令中包含目标音色;

所述通过所述第三转换模型对所述第二内容特征进行音频转换,得到第二口音音频,包括:

将所述第二内容特征以及所述目标音色对应说话者的说话者标识输入所述第三转换模型,得到所述第二口音音频,其中,不同说话者对应不同说话者标识。

12.一种语音转换模型的训练装置,其特征在于,所述装置包括:

训练模块,用于基于第一样本音频训练第一ASR模型,以及基于第二样本音频训练第二

ASR模型，所述第一样本音频对应第一口音，所述第二样本音频对应第二口音；

所述训练模块，还用于基于所述第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，所述第一样本内容特征由所述第一ASR模型对所述第一样本音频进行提取得到，所述第一转换模型用于将文本转换为所述第一口音的内容特征；

所述训练模块，还用于基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，所述第二样本内容特征由所述第二ASR模型对所述第二样本音频进行提取得到，所述平行样本数据由不同内容特征构成，不同内容特征对应不同口音，且不同内容特征对应相同文本；基于所述平行样本数据训练第二转换模型，所述第二转换模型用于对所述第一口音和所述第二口音间进行内容特征转换；

所述训练模块，还用于基于不同样本音频的样本内容特征训练第三转换模型，所述第三转换模型用于将内容特征转换为音频；

生成模块，用于基于训练得到的所述第一ASR模型、所述第二转换模型和所述第三转换模型生成语音转换模型，所述语音转换模型用于将第一口音的音频转换为第二口音的音频。

13.一种计算机设备，其特征在于，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行以实现如权利要求1至11任一所述的语音转换模型的训练方法。

14.一种计算机可读存储介质，其特征在于，所述可读存储介质中存储有至少一条指令，所述至少一条指令由处理器加载并执行以实现如权利要求1至11任一所述的语音转换模型的训练方法。

15.一种计算机程序产品，其特征在于，所述计算机程序产品包括计算机指令，所述计算机指令存储在计算机可读存储介质中；计算机设备的处理器从所述计算机可读存储介质读取所述计算机指令，所述处理器执行所述计算机指令，使得所述计算机设备执行如权利要求1至11任一所述的语音转换模型的训练方法。

语音转换模型的训练方法、装置、设备及介质

技术领域

[0001] 本申请实施例涉及音频处理技术领域，特别涉及一种语音转换模型的训练方法、装置、设备及介质。

背景技术

[0002] 随着网络技术的不断发展，越来越多用户开始使用虚拟形象在网络中进行直播、游戏、社交或者在线会议。

[0003] 为了保护个人隐私安全，用户在使用虚拟形象过程中，可以设置虚拟形象的口音，使原始口音的用户语音被转换为所设置口音后播放，并保证用户语音内容保持不变。相关技术中，口音转换通常使用语音转换模型实现，而在训练语音转换模型过程中，需要基于大量平行语料。其中，该平行语料为相同语音内容的不同口音音频。

[0004] 然而，平行语料通常需要人工录制，导致平行语料的获取难度较高，在平行语料不足的情况下，训练得到语音转换模型的质量较差，进而影响口音转换效果。

发明内容

[0005] 本申请实施例提供了一种语音转换模型的训练方法、装置、设备及介质，能够在降低对人工录制的平行语料的需求的前提下，保证语音转换模型的训练质量。所述技术方案如下：

[0006] 一方面，本申请实施例提供了一种语音转换模型的训练方法，包括：

[0007] 基于第一样本音频训练第一ASR(Automatic Speech Recognition，自动语音识别)模型，以及基于第二样本音频训练第二ASR模型，所述第一样本音频对应第一口音，所述第二样本音频对应第二口音；

[0008] 基于所述第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，所述第一样本内容特征由所述第一ASR模型对所述第一样本音频进行提取得到，所述第一转换模型用于将文本转换为所述第一口音的内容特征；

[0009] 基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，所述第二样本内容特征由所述第二ASR模型对所述第二样本音频进行提取得到，所述平行样本数据由不同内容特征构成，不同内容特征对应不同口音，且不同内容特征对应相同文本；

[0010] 基于所述平行样本数据训练第二转换模型，所述第二转换模型用于对所述第一口音和所述第二口音间进行内容特征转换；

[0011] 基于不同样本音频的样本内容特征训练第三转换模型，所述第三转换模型用于将内容特征转换为音频；

[0012] 基于训练得到的所述第一ASR模型、所述第二转换模型和所述第三转换模型生成语音转换模型，所述语音转换模型用于将第一口音的音频转换为第二口音的音频。

[0013] 另一方面，本申请实施例提供了一种语音转换模型的训练装置，所述装置包括：

[0014] 训练模块，用于基于第一样本音频训练第一ASR模型，以及基于第二样本音频训练第二ASR模型，所述第一样本音频对应第一口音，所述第二样本音频对应第二口音；

[0015] 所述训练模块，还用于基于所述第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，所述第一样本内容特征由所述第一ASR模型对所述第一样本音频进行提取得到，所述第一转换模型用于将文本转换为所述第一口音的内容特征；

[0016] 所述训练模块，还用于基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，所述第二样本内容特征由所述第二ASR模型对所述第二样本音频进行提取得到，所述平行样本数据由不同内容特征构成，不同内容特征对应不同口音，且不同内容特征对应相同文本；基于所述平行样本数据训练第二转换模型，所述第二转换模型用于对所述第一口音和所述第二口音间进行内容特征转换；

[0017] 所述训练模块，还用于基于不同样本音频的样本内容特征训练第三转换模型，所述第三转换模型用于将内容特征转换为音频；

[0018] 生成模块，用于基于训练得到的所述第一ASR模型、所述第二转换模型和所述第三转换模型生成语音转换模型，所述语音转换模型用于将第一口音的音频转换为第二口音的音频。

[0019] 另一方面，本申请实施例提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行以实现如上述方面所述的语音转换模型的训练方法。

[0020] 另一方面，本申请实施例提供了一种计算机可读存储介质，所述可读存储介质中存储有至少一条指令，所述至少一条指令由处理器加载并执行以实现如上述方面所述的语音转换模型的训练方法。

[0021] 另一方面，本申请实施例提供了一种计算机程序产品，所述计算机程序产品包括计算机指令，所述计算机指令存储在计算机可读存储介质中；计算机设备的处理器从所述计算机可读存储介质读取所述计算机指令，所述处理器执行所述计算机指令，使得所述计算机设备执行如上述方面所述的语音转换模型的训练方法。

[0022] 本申请实施例中，在缺少第二口音的第二样本音频对应平行语料的情况下，首先基于第一口音的第一样本音频，训练用于将文本转换为内容特征的第一转换模型，从而利用该第一转换模型以及第二样本音频对应的第二样本文本，构建得到包含对应相同文本内容但对应不同口音的平行样本数据，进而利用该平行样本数据训练在不同口音间进行内容特征转换的第二转换模型，以及用于将内容特征转换为音频的第三转换模型，完成语音转换模型训练；模型训练过程中，利用训练得到的中间模型构建平行语料，无需在模型训练前录制不同口音的平行语料，能够在保证模型训练质量的情况下，降低模型训练对人工录制的平行语料的需求，有助于提高模型训练效率，并提高样本不足情况下模型的训练质量。

附图说明

[0023] 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，对于本领域普通技术人员来讲，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。

[0024] 图1示出了本申请一个示例性实施例提供的实施环境的示意图；

[0025] 图2示出了本申请一个示例性实施例提供的语音转换模型的训练方法的流程图；

[0026] 图3示出了本申请一个示例性实施例提供的口音转换方法的流程图；

[0027] 图4是本申请一个示例性实施例示出的语音设置界面的示意图；

[0028] 图5是本申请一个示例性实施例提供的口音转换过程的实施示意图；

[0029] 图6是本申请一个示例性实施例示出的文本转内容特征过程的流程图；

[0030] 图7是本申请一个示例性实施例提供的FFT结构图；

[0031] 图8是本申请一个示例性实施例示出的第一转换模型的结构示意图；

[0032] 图9是本申请一个示例性实施例示出的第二转换模型训练过程的流程图；

[0033] 图10是本申请一个示例性实施例示出的第二转换模型的结构示意图；

[0034] 图11是本申请一个示例性实施例示出的第三转换模型的结构示意图；

[0035] 图12是本申请一个示例性实施例示出的第三转换模型训练过程的流程图；

[0036] 图13是本申请另一个示例性实施例提供的口音转换过程的实施示意图；

[0037] 图14是本申请一个示例性实施例提供的语音转换模型的训练装置的结构框图；

[0038] 图15示出了本申请一个示例性实施例提供的计算机设备的结构示意图。

具体实施方式

[0039] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

[0040] 相关技术中，采样端到端方式训练语音转换模型时，需要以不同口音朗读同一内容的语音数据作为平行语料。比如，当需要训练将A口音转换为B口音的语音转换模型时，需要获取以A口音以及B口音朗读相同内容的语音数据；当需要训练将A口音转换为C口音的语音转换模型时，需要获取以A口音以及C口音朗读相同内容的语音数据。并且，为了保证训练质量，需要预先准备大量平行语料用于模型训练。

[0041] 显然，相关技术中语音转换模型的训练依赖人工录制的语音，需要花费大量时间准备，在平行语料数量不足的情况下，模型训练质量较差，进而影响口音转换质量。

[0042] 为了降低模型训练过程对预先录制的平行语料的依赖，本申请实施例中，语音转换模型由第一ASR模型(用于将音频转换为文本)、第二转换模型(用于在不同口音间进行内容特征转换)以及第三转换模型(用于将内容特征转换为音频)构成。并且，在训练过程中，完成第一ASR模型训练后，训练用于将文本转换为内容特征的第一转换模型，从而借助第一转换模型构建平行样本数据，以用于后续第二转换模型以及第三转换模型训练。训练过程中，借助训练得到的转换模型构建平行语料，无需预先人工录制大量平行语料，从而降低训练过程对平行语料的依赖，保证模型训练质量。

[0043] 采用本申请实施例提供的训练方法所训练得到的语音转换模型，能够适用于各种需要进行口音转换的场景。如图1所示，其示出了本申请一个示例性实施例示出的实施环境的示意图。该实施环境中包括：音频采集设备110、终端120以及服务器130。

[0044] 音频采集设备110是用于采集用户语音的设备，该音频采集设备110可以是耳麦、麦克风或者具有收音功能的AR/VR设备等等，本申请实施例对此不作限定。

[0045] 音频采集设备110与终端120之间通过有线或无线方式相连，用于将采集到的用户

语音传输至终端120，由终端120进一步对用户语音进行口音转换处理。其中，终端120可以是具有智能手机、平板电脑、个人计算机、车载终端等电子设备。

[0046] 在一些实施例中，终端120中设置有具有口音转换功能的应用程序。通过该应用程序，用户可以设置口音转换目标，从而实现将用户语音由原始语音转换为目标语音。

[0047] 在一种可能的实施方式中，口音转换可以由终端120在本地实现(语音转换模型设置在终端120中)；在另一种可能的实施方式中，口音转换可以由终端120借助服务器130实现(语音转换模型设置在服务器130中，终端120向服务器130传输口音转换需求)。

[0048] 服务器130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、内容分发网络(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。本申请实施例中，服务器130可以为实现口音转换功能的后台服务器，用于提供不同口音间的转换服务。

[0049] 在一些实施例中，服务器130中设置有多个语音转换模型，不同语音转换模型用于实现不同口音间的转换。比如，当支持将普通话转换为n种地方口音时，服务器130中设置有n个语音转换模型。

[0050] 并且，在实现口音转换功能前，服务器130获取不同口音的口音语料，该口音语料由音频以及对应的文本构成，从而基于口音语料训练相应的语音转换模型。

[0051] 如图1所示，在进行口音转换前，用户通过终端120设置将第一口音转换为第二口音，由终端120向服务器130发送口音转换请求，请求服务器130采用相应的语音转换模型(将第一口音转换为第二口音)进行口音转换。

[0052] 音频采集设备110将采集到的第一口音的用户语音传输至终端120，由终端120将第一口音的用户语音传输至服务器130。服务器130通过语音转换模型将其转换为第二口音的用户语音，并反馈至终端120，由终端120进行进一步处理。

[0053] 其中，不同应用场景下，终端120对用户语音的处理方式不同。下面采用集中示例性的应用场景进行说明。

[0054] 1、虚拟人内容制作场景

[0055] 虚拟人内容制作场景下，终端获取转换得到的用户语音后，将该用户语音与制作的内容(比如虚拟人短视频、虚拟人长视频等等)进行融合，得到虚拟人内容。其中，在进行融合时，可以根据转换得到的用户语音对虚拟人的嘴部进行控制，提高虚拟人嘴部动作与语音的匹配度。

[0056] 2、虚拟主播直播场景

[0057] 虚拟主播直播场景下，虚拟主播可以通过口音设置界面预先设置直播口音。直播过程中，终端将通过麦克风采集的用户语音发送至服务器，由服务器将原始口音的用户语音转换为直播口音的用户语音，并反馈至终端。终端将直播口音的用户语音与包含虚拟主播形象的视频流进行合并，从而通过推流服务器将合并得到的音视频流推送至直播间内的各个观众客户端。

[0058] 3、元宇宙场景

[0059] 元宇宙场景下，用户可以设置在元宇宙中进行交互时采用的口音。用户控制虚拟角色在元宇宙中与其他虚拟角色进行交互时，用户语音由耳麦、AR/VR等设备采集并传输至

终端,由终端进一步交由服务器进行口音转换,并在元宇宙中控制虚拟角色播放转换得到的口音音频,实现与其他虚拟角色之间的语音互动。

[0060] 上述应用场景仅作为示例性的说明,采用本申请实施例提供的方法训练得到的语音转换模型,还可以用于语音通话(方便不同口音的通话对象之间进行语音交流)、翻译等真实世界应用场景,本申请实施例并不对此构成限定。

[0061] 并且,为了方便表述,下述各个实施例中,以语音转换模型的训练以及使用均用于计算机设备(可以为终端或者服务器),且训练用于将第一口音转换为第二口音的语音转换模型为例进行说明(其他由源语音转换为目标语音的方案类似),但并不对此构成限定。

[0062] 图2示出了本申请一个示例性实施例提供的语音转换模型的训练方法的流程图。该方法包括如下步骤。

[0063] 步骤201,基于第一样本音频训练第一ASR模型,以及基于第二样本音频训练第二ASR模型,第一样本音频对应第一口音,第二样本音频对应第二口音。

[0064] 其中,第一口音为源口音,第二口音为目标语音,即训练得到的语音转换模型用于将第一口音的语音转换为第二口音的语音。

[0065] 在一些实施例中,第一样本音频对应有第一样本文本,第二样本音频对应有第二样本文本。不同于相关技术中,训练前需要录制对应相同文本但对应不同口音的平行语料,本申请实施例中,第一样本文本无需与第二样本文本相同,因此可以直接采用公开语音数据集用于模型训练。

[0066] 在一个示意性的例子中,计算机设备采用Wenet Speech数据集作为第一样本音频,采用KeSpeech数据集作为第二样本音频,其中,Wenet Speech数据集包括1万小时的ASR数据,KeSpeech数据集包含不同地区方言的ASR数据。

[0067] 关于ASR模型的训练方式,在一种可能的实施方式中,计算机设备将样本音频输入ASR模型,得到ASR模型输出的预测文本,从而基于预测文本以及样本音频对应的样本文本对ASR模型进行训练。

[0068] 可选的,ASR模型的模型架构包括但不限于Wenet、wav2vec2、Kaldi等等,本申请实施例对此不作限定。

[0069] 在一些实施例中,ASR模型可以基于样本音频重新训练得到(适用于样本音频数量较多的情况),也可以基于样本音频对预训练的ASR模型进行微调得到(适用于样本音频数量较少的情况)。

[0070] 比如,当第一口音为普通话,第二口音为方言时,第一ASR模型基于第一样本音频重新训练得到,第二ASR模型在第一ASR模型的基础上,基于第二样本音频进行微调得到。

[0071] 本申请实施例中,训练得到ASR模型用于提取语音中的内容特征。在一些实施例中,该内容特征被称为BN(BottleNeck,瓶颈)特征,通常为ASR模型的最后一层特征,其保留了语音的内容特征,并剔除了诸如音色、音调等其他特征。

[0072] 步骤202,基于第一样本音频对应的第一样本文本以及第一样本内容特征,训练第一转换模型,第一样本内容特征由第一ASR模型对第一样本音频进行提取得到,第一转换模型用于将文本转换为第一口音的内容特征。

[0073] 由于采用不同口音表述相同文本内容时的发音并不相同,因此对相同文本对应的不同口音语音进行内容特征提取所得到的内容特征也不同,相应的,在不同口音间实现内

容特征转换成为实现口音转换的关键。

[0074] 不同于相关技术中,需要使用相同文本对应的不同口音语音作为平行语料训练转换模块,本申请实施例中采用了数据增强方案实现非平行语料(即对应不同口音且对应不同文本的语料)之间的内容特征转换。

[0075] 在一些实施例中,计算机设备通过训练得到的第一ASR模型对第一样本音频进行特征提取,得到第一样本音频的第一样本内容特征,从而基于第一样本音频对应的第一样本文本以及第一样本内容特征,训练第一转换模型。其中,第一转换模型可以被称为文本内容特征转换模型(Text2BN模型),用于实现文本至源口音内容特征之间的转换。

[0076] 步骤203,基于第一转换模型、第二样本音频对应的第二样本文本以及第二样本内容特征,构建平行样本数据,第二样本内容特征由第二ASR模型对第二样本音频进行提取得到,平行样本数据由不同内容特征构成,不同内容特征对应不同口音,且不同内容特征对应相同文本。

[0077] 训练得到第一转换模型后,计算机设备基于第二样本音频对应的第二样本文本以及第一转换模型进行数据增强,从而根据第二样本内容特征以及数据增强得到的第一口音的内容特征,构建得到平行样本数据。其中,平行样本数据由相同文本对应的第一口音的内容特征(由第一转换模型生成)以及第二口音的内容特征(由第二ASR模型提取得到)。

[0078] 比如,在包含对应文本A的方言样本音频,而不包含对应文本A的普通话样本音频的情况下,计算机设备可以基于第一转换模型、文本A以及对应文本A的方言样本音频的方言样本内容特征,构建得到文本A对应的平行样本数据,该平行样本数据包含文本A对应的普通话以及方言的内容特征。

[0079] 步骤204,基于平行样本数据训练第二转换模型,第二转换模型用于对第一口音和第二口音间进行内容特征转换。

[0080] 进一步的,计算机设备基于对应相同文本的平行样本数据,训练第二转换模型。其中,该第二转换模型可以被称为内容特征转换模型(BN2BN模型),用于将源口音的内容特征转换为目标口音的内容特征。

[0081] 比如,当第一口音为普通话,第二口音为方言时,计算机设备训练得到第二转换模型用于将普通话的内容特征转换为方言的内容特征。

[0082] 需要说明的是,当第一样本音频与第二样本音频对应相同样本文本时,第一样本音频以及第二样本音频对应的样本内容特征可以直接被用于训练第二转换模型。

[0083] 步骤205,基于不同样本音频的样本内容特征训练第三转换模型,第三转换模型用于将内容特征转换为音频。

[0084] 其中,第三转换模型可以被称为内容音频转换模型(BN2Wav模型),用于基于目标语音的内容特征转换得到目标语音的音频。

[0085] 在一些实施例中,该第三转换模型可以由声学模型以及声码器构成,其中,声学模型用于基于内容特征生成音频频谱,而声码器则用于基于音频频谱生成音频。

[0086] 在一些实施例中,训练第三转换模型的样本可以为各种口音的样本音频。

[0087] 需要说明的是,第三转换模型可以在ASR模型训练完成后执行,即第三转换模型可以与第一以及第二转换模型同步训练。本申请实施例并不对模型的训练时序构成限定。

[0088] 步骤206,基于训练得到的第一ASR模型、第二转换模型和第三转换模型生成语音

转换模型,语音转换模型用于将第一口音的音频转换为第二口音的音频。

[0089] 通过上述步骤训练得到第一ASR模型、第二转换模型和第三转换模型后,计算机设备将上述模型组合得到最终的语音转换模型。其中,模型之间的拼接顺序为第一ASR模型→第二转换模型→第三转换模型,即第一ASR模型的输出被输入第二转换模型,第二转换模型的输出被输入第三转换模型。

[0090] 在一个示意性的例子中,训练得到的用于将普通话转换为方言的语音转换模型由普通话ASR模型,普通话-方言内容转换模型以及内容音频转换模型构成。

[0091] 综上所述,本申请实施例中,在缺少第二口音的第二样本音频对应平行语料的情况下,首先基于第一口音的第一样本音频,训练用于将文本转换为内容特征的第一转换模型,从而利用该第一转换模型以及第二样本音频对应的第二样本文本,构建得到包含对应相同文本内容但对应不同口音的平行样本数据,进而利用该平行样本数据训练在不同口音间进行内容特征转换的第二转换模型,以及用于将内容特征转换为音频的第三转换模型,完成语音转换模型训练;模型训练过程中,利用训练得到的中间模型构建平行语料,无需在模型训练前录制不同口音的平行语料,能够在保证模型训练质量的情况下,降低模型训练对人工录制的平行语料的需求,有助于提高模型训练效率,并提高样本不足情况下模型的训练质量。

[0092] 下面对采用上述方案训练得到的语音转换模型的应用过程进行说明。图3示出了本申请一个示例性实施例提供的口音转换方法的流程图。该方法包括如下步骤。

[0093] 步骤301,响应于口音转换指令,通过第一ASR模型提取第一口音音频的第一内容特征,第一内容特征对应第一口音,口音转换指令用于指示将音频由第一口音转换为第二口音。

[0094] 在一些实施例中,该口音转换指令在完成口音设置后触发。在一种可能的场景中,如图4所示,在元宇宙虚拟角色设置界面41中,除了包含虚拟角色形象设置选项外,还包括语音设置选项。用户可以通过语音设置选项设置虚拟角色的音色以及口音。当完成语音以及形象设置后,通过触发进入按键42即可进入元宇宙。其中,触发进入按键42后,计算机设备接收到口音转换指令,该口音转换指令中包含源口音以及目标口音的口音标识。本实施例中,以源口音为第一口音,目标口音为第二口音为例进行说明。

[0095] 计算机设备接收到第一口音的第一口音音频后,通过语音转换模型中的第一ASR模型进行内容特征提取,得到第一内容特征,该第一内容特征提出了音色、音调等干扰,仅保留所表达内容层面的特征。

[0096] 在一些实施例中,计算机设备将第一ASR模型最后一层BN特征作为第一内容特征。

[0097] 示意性的,如图5所示,当需要将普通话转换为方言时,计算机设备通过普通话ASR模型52对普通话音频51进行特征提取,得到普通话内容特征53。

[0098] 步骤302,通过第二转换模型将第一内容特征转换为第二内容特征,第二内容特征对应第二口音。

[0099] 进一步的,计算机设备将第一ASR模型提取到的第一内容特征输入第二转换模型,由第二转换模型在第一口音和第二口音之间进行内容特征转换,得到第二口音下的第二内容特征。其中,第一内容特征和第二内容特征对应相同文本(均为第一口音音频对应的文本)。

[0100] 示意性的，如图5所示，BN2BN模型54用于在普通话和方言之间进行内容特征转换。得到普通话内容特征53后，计算机设备进一步通过BN2BN模型54对普通话内容特征53进行特征转换，得到方言内容特征55。

[0101] 步骤303，通过第三转换模型对第二内容特征进行音频转换，得到第二口音音频。

[0102] 进一步的，计算机设备将第二内容特征输入第三转换模型，由第三转换模型基于内容特征生成第二口音音频。

[0103] 示意性的，如图5所示，计算机设备将方言内容特征55输入BN2Wav模型56，得到BN2Wav模型56输出的方言音频57。

[0104] 第一转换模型作为构建平行样本数据的关键模型，在训练第一转换模型的过程中，计算机设备将第一样本文本输入第一转换模型，得到第一转换模型输出的第一预测内容特征，从而以第一样本内容特征为第一预测内容特征的监督，训练第一转换模型。

[0105] 在一些实施例中，计算机设备以第一样本内容特征为第一预测内容特征的监督，基于第一预测内容特征与第一样本内容特征之间的特征差异确定损失，从而基于该损失训练第一转换模型。其中，该损失可以为MSE(Mean Square Error，均方误差)损失或其他类型损失，本实施例对此不作限定。

[0106] 可选的，第一转换模型的损失$loss_{Text2BN}$可以表示为：

[0107] $$loss_{Text2BN} = \left\| BN_{na} - \widehat{BN}_{na} \right\|$$

[0108] 其中，BN_{na}为第一ASR模型提取到的第一样本内容特征，$\widehat{BN}_{na}$为第一转换模型输出的第一预测内容特征。

[0109] 为了提升文本到内容特征的转换质量，在一种可能的设计中，该第一转换模型由第一转换子模型、时长预测子模型以及第二转换子模型构成，其中，第一转换子模型用于实现文本到文本编码特征之间的转换，时长预测子模型用于预测文本的表述时长，而第二转换子模型则用于将文本编码特征转换为内容特征。

[0110] 相应的，第一转换模型将文本转换为内容特征的过程如图6所示。

[0111] 步骤601，通过第一转换子模型对第一样本文本进行编码，得到第一文本编码特征。

[0112] 由于文本表述具有前后关联性，因此本申请实施例中，为了提高后续特征转换质量，在一种可能的设计中，采用N层堆叠的FFT(Feed Forward Transformer，前馈转换)构成第一转换子模型。其中，该FFT用于通过线性变换，先将数据映射到高纬度的空间再映射到低纬度的空间，以此提取更深层次的特征。

[0113] 并且，FFT由多头注意力机制层和卷积层构成。在一个示意性的例子中，该FFT结构如图7所示。原始输入首先经过多头注意力层处理，多头注意力处理得到的多路结果和原始输入共同经过加权和标准化处理后，输入卷积层进行卷积处理。卷积层的输入与输出相加后继续进行加权和标准化处理最终输出。

[0114] 由于FFT通过多头注意力机制和卷积层实现，且使用了残差网络思想，因此利用多层FFT叠加得到的第一转换子模型进行文本编码，能够提高文本编码质量。

[0115] 当然，第一转换子模型除了可以采用堆叠的FFT外，还可以采用LSTM(Long Short-Term Memory，长短期记忆)等其他类型的模块(需要包含注意力机制，且保持输入与输出的

尺寸一致)实现,本申请实施例对此不作限定。

[0116] 步骤602,通过时长预测子模型对第一文本编码特征进行时长预测,得到预测时长,预测时长用于表征第一样本文本的发音时长。

[0117] 由于采用口语表述文本时,具有一定的表述时长,因此为了提高后续转换得到的音频的真实性(使转换的得到的语音符合真人语速),计算机设备通过时长预测子模型进行时长预测,得到第一样本文本的发音时长。

[0118] 在一些实施例中,该预测时长包括第一样本文本中各个子文本对应的发音子时长。比如,第一样本文本为“今天天气真好”,该预测时长包括“今”、“天”、“天”、“气”、“真”、“好”各自对应的发音时长。

[0119] 步骤603,基于预测时长对第一文本编码特征进行特征扩充,得到第二文本编码特征。

[0120] 进一步的,计算机设备基于预测时长对第一文本编码特征进行特征扩充,复制第一文本编码特征中的子特征,使复制后子特征对应的持续时长与对应子文本的发音子时长保持一致。

[0121] 在一个示意性的例子中,第一文本编码特征为“abcd”,经过特征扩充后的第二文本编码特征为“aabbbcdddd”。

[0122] 步骤604,通过第二转换子模型对第二文本编码特征进行转换,得到第一预测内容特征。

[0123] 在一些实施例中,第二转换子模型输出的第一预测内容特征与输出的第二文本编码特征的特征尺寸保持一致。

[0124] 在一些实施例中,第二转换子模型由N层FFT堆叠形成,以此提升文本编码特征到内容特征的转换质量。

[0125] 在一个示意性的例子中,如图8所示,第一转换子模型81首先对第一样本文本进行特征编码,得到第一文本编码特征,并将第一文本编码特征输入时长预测子模型82,得到预测时长,并基于预测时长对第一文本编码特征进行特征扩充处理,得到第二文本编码特征。最终通过第二转换子模型83对第二文本编码特征进行特征转换,得到第一预测内容特征。

[0126] 针对上述平行样本数据的构建,以及基于平行样本数据训练第二转换模型的过程,在一种可能的实施方式中,如图9所示,该过程可以包括如下步骤。

[0127] 步骤901,通过第一转换模型对第二样本文本进行转换,得到第三样本内容特征,第三样本内容特征指采用第一口音表述第二样本文本所产生音频的内容特征。

[0128] 在基于第二样本音频构建平行样本数据时,计算机设备对第二样本音频对应的第二样本文本进行内容特征转换,得到第三样本内容特征。由于第一转换模型用于将文本转换为第一口音的内容特征,因此利用第一转换模型对第二样本文本进行内容特征转换,得到第三样本内容特征即为采用第一口音表述第二样本文本所产生音频的内容特征。

[0129] 借助第一转换模型,即便缺少第二样本音频对应的平行语料,也能够生成平行语料的内容特征,免去了人工录制平行语料,以及对平行语料进行内容特征提取的流程。

[0130] 步骤902,基于第二样本内容特征和第三样本内容特征构建平行样本数据。

[0131] 由于第三样本内容特征与第二样本内容特征对应不同口音,且对应相同文本,因此两者组合集合构建得到平行样本数据。

[0132] 步骤903,将第三样本内容特征输入第二转换模型,得到第二预测内容特征。

[0133] 在一种可能的设计中,为了提高内容特征转换质量,第二转换模型由卷积层以及N层堆叠的FFT构成,其中,FFT的具体结构可以参考图7,本实施例在此不做赘述。进行内容特征转换时,内容特征首先经过卷积层卷积处理,然后经过N层FFT处理,得到转换后的内容特征。

[0134] 示意性的,如图10所示,计算机设备通过卷积层1001对第三样本内容特征进行卷积处理后,将卷积结果输入N层FFT1002,得到第二预测内容特征。

[0135] 步骤904,以第二样本内容特征为第二预测内容特征的监督,训练第二转换模型。

[0136] 为了使第二转换模型的内容转换结果接近第二ASR模型输出的第二样本音频的第二样本内容特征,在一种可能的实施方式中,计算机设备基于第二样本内容特征与第二预测内容特征的差异确定损失,从而基于该损失训练第二转换模型。

[0137] 其中,该损失可以为MSE损失或其他类型损失,本实施例对此不作限定。

[0138] 可选的,第二转换模型的损失$loss_{BN2BN}$可以表示为:

[0139] $$loss_{BN2BN} = \left\| BN_{ac} - \widehat{BN}_{ac} \right\|$$

[0140] 其中,BN_{ac}为第二ASR模型提取到的第二样本内容特征,$\widehat{BN}_{ac}$为第二转换模型输出的第二预测内容特征。

[0141] 由于内容特征剔除了音色等因素的影响,而样本音频具有音色特征,因此在训练第三转换模型过程中,需要将样本音频的说话者标识作为输入的一部分,使训练得到的第三转换模型能够输出具有特定音色的音频。

[0142] 在一种可能的实施方式中,计算机设备将样本内容特征以及样本音频对应的说话者标识输入第三转换模型,得到预测音频,从而基于预测音频以及样本音频,训练第三转换模型。其中,预测音频与样本音频对应相同音频内容,且具有相同音色。

[0143] 可选的,不同说话者对应不同说话者标识。在一些实施例中,预先将说话者划分为不同音色,从而为同一音色对应的不同说话者分配相同说话者标识。

[0144] 在一种可能的设计中,第三转换模型由第三转换子模型以及声码器构成,其中,第三转换子模型用于将内容特征转换为音频谱特征,声码器则用于基于音频谱特征生成音频。

[0145] 可选的,该第三转换模型可以由卷积层以及N层堆叠的FFT构成,该音频谱特征可以为Mel(梅尔)谱特征、MFCC(Mel Frequency Cepstrum Coefficient,梅尔频率倒谱系数)特征等等,本申请实施例对此不作限定。

[0146] 可选的,该声码器可以是采用自回归的Wavenet或WaveRNN,或者采用非自回归的hifigan或melgan等等,本申请实施例对此不作限定。

[0147] 为了方便表述,下述实施例中以音频谱特征为Mel谱特征,声码器为hifigan为例进行说明,但并不对此构成限定。

[0148] 相应的,训练过程中,计算机设备将样本内容特征以及说话者标识输入第三转换子模型,得到预测音频谱特征,从而将预测音频谱特征输入声码器,得到预测音频。

[0149] 示意性的,如图11所示,BN2Wav模型由BN2Mel子模型1101以及hifigan子模型1102构成,其中,BN2Mel子模型1101由卷积层11011以及N层堆叠的FFT 11012构成。模型训练过

程中，计算机设备将样本音频的样本内容特征BN以及说话者标识spk_id输入BN2Mel子模型1101。BN2Mel子模型1101将转换得到的Mel频谱输入hifigan子模型1102，由hifigan子模型1102转换得到预测音频。

[0150] 在一种可能的实施方式中，计算机设备联合训练第三转换子模型和声码器。

[0151] 在另一种可能的实施方式中，计算机设备首先训练第三转换子模型，然后基于训练完成的第三转换子模型训练声码器，以此提高训练效率。

[0152] 如图12所示，第三转换模型的训练过程可以包括如下步骤。

[0153] 步骤1201，将样本内容特征以及说话者标识输入第三转换子模型，得到预测音频谱特征。

[0154] 在一种可能的实施方式中，计算机设备将样本内容特征以及说话者标识输入第三转换子模型，得到样本音频对应的预测Mel频谱。

[0155] 步骤1202，以样本音频的样本音频谱特征为预测音频谱特征的监督，训练第三转换子模型。

[0156] 在一些实施例中，计算机设备对样本音频进行音频谱特征提取，得到样本音频谱特征，从而基于预测音频谱特征与样本音频谱特征的差异确定损失，从而基于该损失训练第三转换子模型。

[0157] 其中，该损失可以为MSE损失或其他类型损失，本实施例对此不作限定。

[0158] 可选的，第三转换子模型的损失$loss_{BN2Mel}$可以表示为：

[0159] $$loss_{BN2Mel} = \|Mel - \widehat{Mel}\|$$

[0160] 其中，Mel为直接对样本音频进行音频谱特征提取到的样本音频谱特征，$\widehat{Mel}$为第三转换子模型输出的预测音频谱特征。

[0161] 步骤1203，在第三转换子模型训练完成的情况下，将训练完成后第三转换子模型输出的预测音频谱特征输入声码器，得到预测音频。

[0162] 完成第三转换子模型训练后，计算机设备将样本内容特征以及说话者标识输入训练得到的第三转化子模型，得到预测音频谱特征，然后将该预测音频谱特征输入声码器，得到声码器输出的预测音频。

[0163] 在一个示意性的例子中，计算机设备将训练完成的BN2Mel子模型输出的预测Mel频谱特征输入hifigan，得到hifigan输出的预测音频。

[0164] 步骤1204，基于预测音频以及样本音频，训练第三转换模型中的声码器。

[0165] 可选的，计算机设备以样本音频为预测音频的监督，确定声码器的转换损失，从而基于该损失训练声码器。

[0166] 在一些实施例中，当声码器采用对抗网络时，以hifigan为例，计算机设备采用对抗训练思想，通过生成器(Generator)和判别器(Discriminator)对抗训练。对抗训练过程中生成器的损失可以表示为：

[0167] $L_G = L_G(G;D) + L_{FM}(G;D) + L_{mel}(G)$

[0168] 其中，$L_{mel}(G) = \|\varphi(x) - \varphi(G(s))\|$，$\varphi(G(s))$为生成器生成的音频G(s)重新转换得到的Mel谱特征，$\varphi(x)$为从样本音频中提取到的Mel谱特征；$L_{FM}(G;D)$为生成音频与样本音频的特征匹配损失；$L_G(G;D)$为生成音频的判别损失。

[0169] 对抗训练过程中判别器的损失可以表示为：

[0170] $L_D(G;D) = (D(x)-1)^2 + (D(G(s)))^2$

[0171] 其中，D(x)为判别器对样本音频的判别结果，D(G(s))为判别器对预测音频的判别结果。

[0172] 显然，通过上述方式训练得到的第三转换模型，处理能够将内容特征转换为音频外，还能够在转换得到的音频中添加特定的音色。相应的，在应用过程中，用户除了选择目标口音外，还可以选择目标音色。

[0173] 在一种可能的实施方式中，在口音转换指令中包含目标音色的情况下，计算机设备将第二内容特征以及目标音色对应说话者的说话者标识输入第三转换模型，得到第二口音音频，其中，第二口音音频具有第二口音以及目标音色。

[0174] 示意性的，如图13所示，当需要将普通话转换为方言，且具有目标音色时，计算机设备通过普通话ASR模型1302对普通话音频1301进行特征提取，得到普通话内容特征1303。计算机设备进一步通过BN2BN模型1304对普通话内容特征1303进行特征转换，得到方言内容特征1305。计算机设备将方言内容特征1305以及目标音色1306对应的说话者标识输入BN2Wav模型1307，得到BN2Wav模型1307输出的带有目标音色的方言音频1308。

[0175] 需要说明的是，当需要保持口音转换前后音色一致时，需要预先获取当前用户的语料数据(比如累计30分钟时长的语音数据)，并为当前用户分配说话者标识，从而基于当前用户的语料数据以及说话者标识训练第三转换模型，本实施例在此不作赘述。

[0176] 本实施例中，在训练第三转换模型的过程中，除了将样本音频的内容特征作为输入外，还将样本音频对应说话者标识作为输入，使第三转换模型在训练中能够基于内容特征以及说话者的音色特征进行音频转换。后续使用过程中，对于相同文本内容，通过输入不同的说话者标识，第三转换模型能够输出不同音色的音频，实现口音以及音色的双重转换。

[0177] 图14是本申请一个示例性实施例提供的语音转换模型的训练装置的结构框图，该装置包括：

[0178] 训练模块1401，用于基于第一样本音频训练第一ASR模型，以及基于第二样本音频训练第二ASR模型，所述第一样本音频对应第一口音，所述第二样本音频对应第二口音；

[0179] 所述训练模块1401，还用于基于所述第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，所述第一样本内容特征由所述第一ASR模型对所述第一样本音频进行提取得到，所述第一转换模型用于将文本转换为所述第一口音的内容特征；

[0180] 所述训练模块1401，还用于基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，所述第二样本内容特征由所述第二ASR模型对所述第二样本音频进行提取得到，所述平行样本数据由不同内容特征构成，不同内容特征对应不同口音，且不同内容特征对应相同文本；基于所述平行样本数据训练第二转换模型，所述第二转换模型用于对所述第一口音和所述第二口音间进行内容特征转换；

[0181] 所述训练模块1401，还用于基于不同样本音频的样本内容特征训练第三转换模型，所述第三转换模型用于将内容特征转换为音频；

[0182] 生成模块1402，用于基于训练得到的所述第一ASR模型、所述第二转换模型和所述第三转换模型生成语音转换模型，所述语音转换模型用于将第一口音的音频转换为第二口

音的音频。

[0183] 可选的，所述训练模块1401，用于：

[0184] 通过所述第一转换模型对所述第二样本文本进行转换，得到第三样本内容特征，所述第三样本内容特征指采用所述第一口音表述第二样本文本所产生音频的内容特征；

[0185] 基于所述第二样本内容特征和所述第三样本内容特征构建所述平行样本数据。

[0186] 可选的，所述训练模块1401，用于：将所述第三样本内容特征输入所述第二转换模型，得到第二预测内容特征；

[0187] 以所述第二样本内容特征为所述第二预测内容特征的监督，训练所述第二转换模型。

[0188] 可选的，所述训练模块1401，用于：将所述第一样本文本输入所述第一转换模型，得到所述第一转换模型输出的第一预测内容特征；

[0189] 以所述第一样本内容特征为所述第一预测内容特征的监督，训练所述第一转换模型。

[0190] 可选的，所述第一转换模型中包括第一转换子模型、时长预测子模型以及第二转换子模型；

[0191] 所述训练模块1401，用于：

[0192] 通过所述第一转换子模型对所述第一样本文本进行编码，得到第一文本编码特征；

[0193] 通过所述时长预测子模型对所述第一文本编码特征进行时长预测，得到预测时长，所述预测时长用于表征所述第一样本文本的发音时长；

[0194] 基于所述预测时长对所述第一文本编码特征进行特征扩充，得到第二文本编码特征；

[0195] 通过所述第二转换子模型对所述第二文本编码特征进行转换，得到所述第一预测内容特征。

[0196] 可选的，所述第一转换子模型和所述第二转换子模型由FFT堆叠而成，所述FFT由多头注意力机制层和卷积层构成。

[0197] 可选的，所述训练模块1401，用于：

[0198] 将所述样本内容特征以及所述样本音频对应的说话者标识输入所述第三转换模型，得到预测音频，所述预测音频与所述样本音频对应相同音频内容，且具有相同音色，其中，不同说话者对应不同说话者标识；

[0199] 基于所述预测音频以及所述样本音频，训练所述第三转换模型。

[0200] 可选的，所述第三转换模型包括第三转换子模型以及声码器；

[0201] 所述训练模块1401，用于：

[0202] 将所述样本内容特征以及所述说话者标识输入所述第三转换子模型，得到预测音频谱特征；

[0203] 将所述预测音频谱特征输入所述声码器，得到所述预测音频。

[0204] 可选的，所述训练模块1401，用于：

[0205] 以所述样本音频的样本音频谱特征为所述预测音频谱特征的监督，训练所述第三转换子模型；

[0206] 在所述第三转换子模型训练完成的情况下,将训练完成后所述第三转换子模型输出的所述预测音频谱特征输入所述声码器,得到所述预测音频;

[0207] 基于所述预测音频以及所述样本音频,训练所述第三转换模型中的所述声码器。

[0208] 可选的,所述装置还包括:

[0209] 转换模块,用于响应于口音转换指令,通过所述第一ASR模型提取第一口音音频的第一内容特征,第一内容特征对应所述第一口音,所述口音转换指令用于指示将音频由所述第一口音转换为所述第二口音;

[0210] 通过所述第二转换模型将所述第一内容特征转换为第二内容特征,所述第二内容特征对应所述第二口音;

[0211] 通过所述第三转换模型对所述第二内容特征进行音频转换,得到第二口音音频。

[0212] 可选的,所述口音转换指令中包含目标音色;

[0213] 所述转换模块,用于:

[0214] 将所述第二内容特征以及所述目标音色对应说话者的说话者标识输入所述第三转换模型,得到所述第二口音音频,其中,不同说话者对应不同说话者标识。

[0215] 综上所述,本申请实施例中,在缺少第二口音的第二样本音频对应平行语料的情况下,首先基于第一口音的第一样本音频,训练用于将文本转换为内容特征的第一转换模型,从而利用该第一转换模型以及第二样本音频对应的第二样本文本,构建得到包含对应相同文本内容但对应不同口音的平行样本数据,进而利用该平行样本数据训练在不同口音间进行内容特征转换的第二转换模型,以及用于将内容特征转换为音频的第三转换模型,完成语音转换模型训练;模型训练过程中,利用训练得到的中间模型构建平行语料,无需在模型训练前录制不同口音的平行语料,能够在保证模型训练质量的情况下,降低模型训练对人工录制的平行语料的需求,有助于提高模型训练效率,并提高样本不足情况下模型的训练质量。

[0216] 需要说明的是:上述实施例提供的装置,仅以上述各功能模块的划分进行举例说明,实际应用中,可以根据需要而将上述功能分配由不同的功能模块完成,即将装置的内部结构划分成不同的功能模块,以完成以上描述的全部或者部分功能。另外,上述实施例提供的装置与方法实施例属于同一构思,其实现过程详见方法实施例,这里不再赘述。

[0217] 请参考图15,其示出了本申请一个示例性实施例提供的计算机设备的结构示意图,该计算机设备可以为上述实施例中的投屏设备或终端。具体来讲:所述计算机设备1600包括中央处理单元(Central Processing Unit,CPU)1601、包括随机存取存储器1602和只读存储器1603的系统存储器1604,以及连接系统存储器1604和中央处理单元1601的系统总线1605。所述计算机设备1600还包括帮助计算机内的各个器件之间传输信息的基本输入/输出系统(Input/Output,I/O系统)1606,和用于存储操作系统1613、应用程序1614和其他程序模块1615的大容量存储设备1607。

[0218] 所述基本输入/输出系统1606包括有用于显示信息的显示器1608和用于用户输入信息的诸如鼠标、键盘之类的输入设备1609。其中所述显示器1608和输入设备1609都通过连接到系统总线1605的输入输出控制器1610连接到中央处理单元1601。所述基本输入/输出系统1606还可以包括输入输出控制器1610以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地,输入输出控制器1610还提供输出到显示屏、打印机或

其他类型的输出设备。

[0219] 所述大容量存储设备1607通过连接到系统总线1605的大容量存储控制器(未示出)连接到中央处理单元1601。所述大容量存储设备1607及其相关联的计算机可读介质为计算机设备1600提供非易失性存储。也就是说,所述大容量存储设备1607可以包括诸如硬盘或者驱动器之类的计算机可读介质(未示出)。

[0220] 不失一般性,所述计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括随机存取记忆体(RAM,Random Access Memory)、只读存储器(ROM,Read Only Memory)、闪存或其他固态存储其技术,只读光盘(Compact Disc Read-Only Memory,CD-ROM)、数字通用光盘(Digital Versatile Disc,DVD)或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然,本领域技术人员可知所述计算机存储介质不局限于上述几种。上述的系统存储器1604和大容量存储设备1607可以统称为存储器。

[0221] 存储器存储有一个或多个程序,一个或多个程序被配置成由一个或多个中央处理单元1601执行,一个或多个程序包含用于实现上述方法的指令,中央处理单元1601执行该一个或多个程序实现上述各个方法实施例提供的方法。

[0222] 根据本申请的各种实施例,所述计算机设备1600还可以通过诸如因特网等网络连接到网络上的远程计算机运行。也即计算机设备1600可以通过连接在所述系统总线1605上的网络接口单元1611连接到网络1612,或者说,也可以使用网络接口单元1611来连接到其他类型的网络或远程计算机系统(未示出)。

[0223] 本申请实施例还提供一种计算机可读存储介质,该可读存储介质中存储有至少一条指令,至少一条指令由处理器加载并执行以实现上述实施例所述的语音转换模型的训练方法。

[0224] 可选地,该计算机可读存储介质可以包括:ROM、RAM、固态硬盘(SSD,Solid State Drives)或光盘等。其中,RAM可以包括电阻式随机存取记忆体(ReRAM,Resistance Random Access Memory)和动态随机存取存储器(DRAM,Dynamic Random Access Memory)。

[0225] 本申请实施例提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令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指令,处理器执行该计算机指令,使得该计算机设备执行上述实施例所述的语音转换模型的训练方法。

[0226] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成,也可以通过程序来指令相关的硬件完成,所述的程序可以存储于一种计算机可读存储介质中,上述提到的存储介质可以是只读存储器,磁盘或光盘等。

[0227] 以上所述仅为本申请的可选的实施例,并不用以限制本申请,凡在本申请的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本申请的保护范围之内。

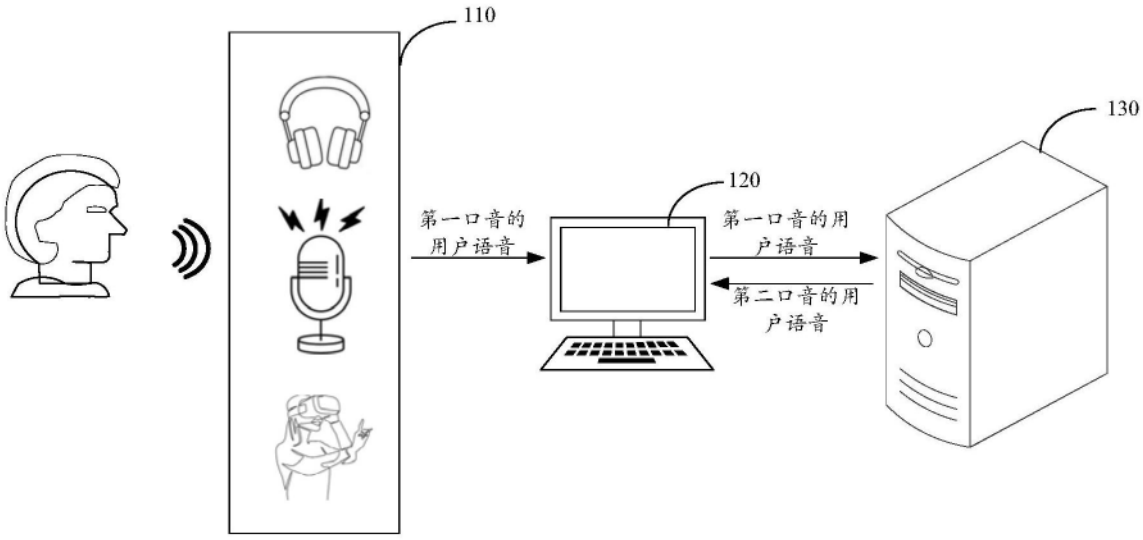

110
130
120
第一口音的
用户语音
第一口音的用
户语音
第二口音的用
户语音

图1

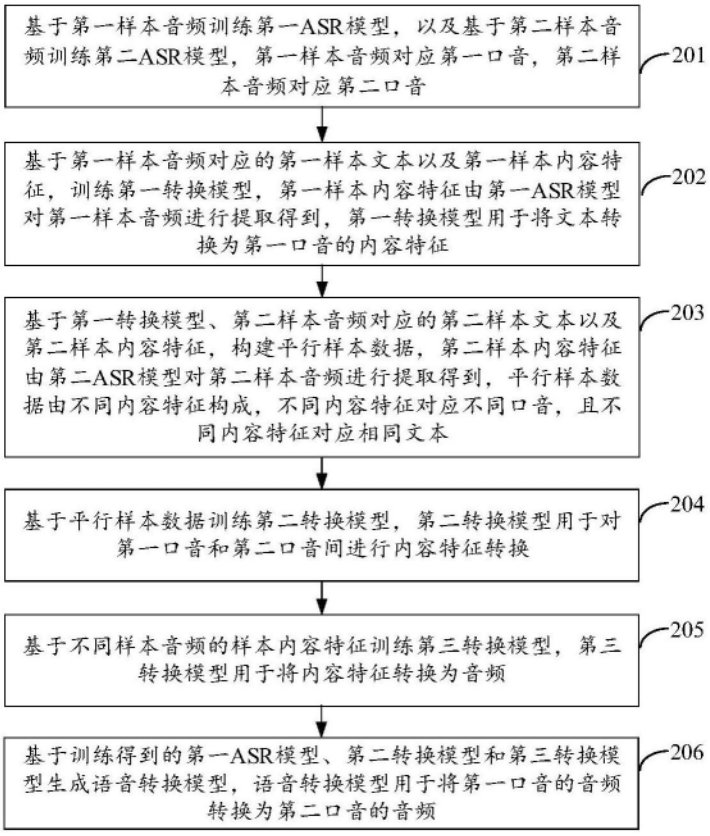

基于第一样本音频训练第一ASR模型，以及基于第二样本音频训练第二ASR模型，第一样本音频对应第一口音，第二样本音频对应第二口音
201
基于第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，第一样本内容特征由第一ASR模型对第一样本音频进行提取得到，第一转换模型用于将文本转换为第一口音的内容特征
202
基于第一转换模型、第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，第二样本内容特征由第二ASR模型对第二样本音频进行提取得到，平行样本数据由不同内容特征构成，不同内容特征对应不同口音，且不同内容特征对应相同文本
203
基于平行样本数据训练第二转换模型，第二转换模型用于对第一口音和第二口音间进行内容特征转换
204
基于不同样本音频的样本内容特征训练第三转换模型，第三转换模型用于将内容特征转换为音频
205
基于训练得到的第一ASR模型、第二转换模型和第三转换模型生成语音转换模型，语音转换模型用于将第一口音的音频转换为第二口音的音频
206

图2

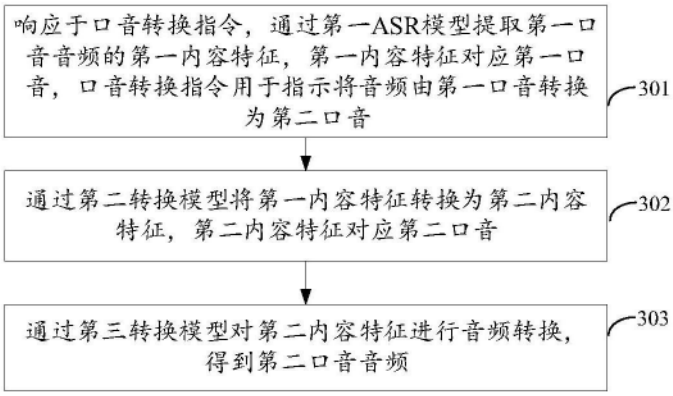

响应于口音转换指令，通过第一ASR模型提取第一口音音频的第一内容特征，第一内容特征对应第一口音，口音转换指令用于指示将音频由第一口音转换为第二口音
301
通过第二转换模型将第一内容特征转换为第二内容特征，第二内容特征对应第二口音
302
通过第三转换模型对第二内容特征进行音频转换，得到第二口音音频
303

图3

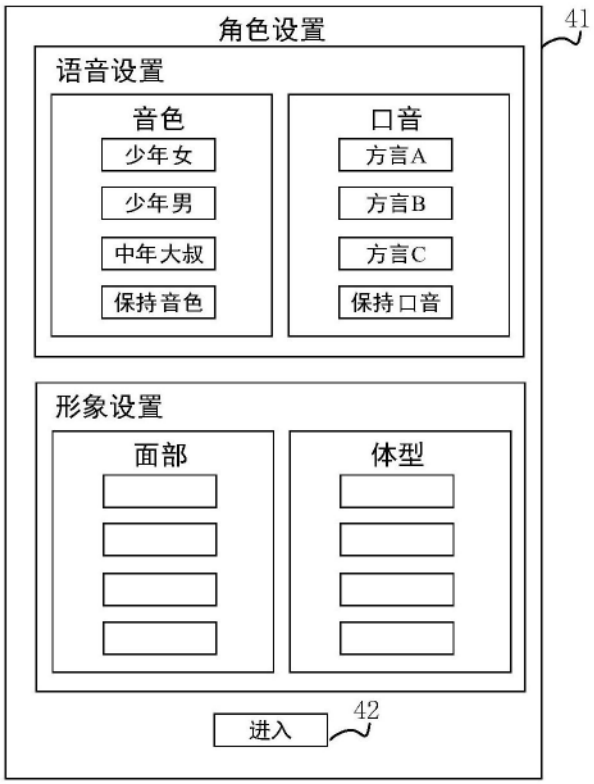

角色设置
41
语音设置
音色
少年女
少年男
中年大叔
保持音色
口音
方言A
方言B
方言C
保持口音
形象设置
面部
体型
进入
42

图4

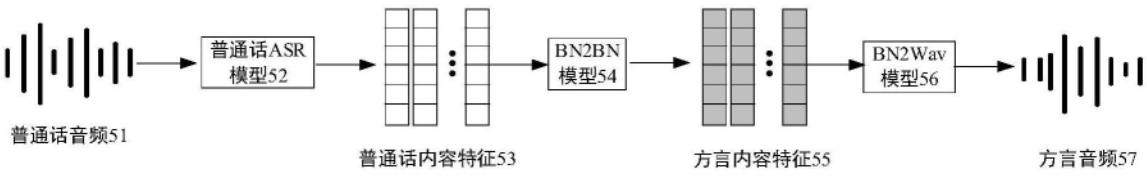

普通话音频51
普通话ASR模型52
普通话内容特征53
BN2BN模型54
方言内容特征55
BN2Wav模型56
方言音频57

图5

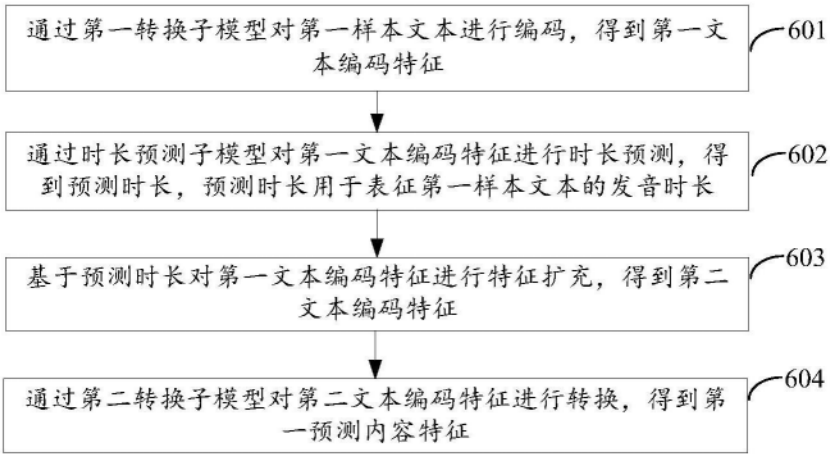

通过第一转换子模型对第一样本文本进行编码，得到第一文本编码特征
601
通过时长预测子模型对第一文本编码特征进行时长预测，得到预测时长，预测时长用于表征第一样本文本的发音时长
602
基于预测时长对第一文本编码特征进行特征扩充，得到第二文本编码特征
603
通过第二转换子模型对第二文本编码特征进行转换，得到第一预测内容特征
604

图6

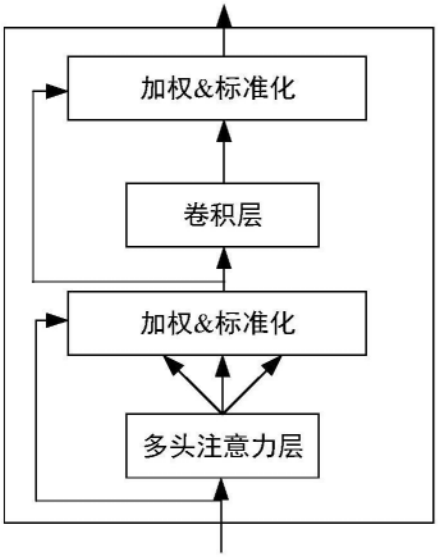

加权&标准化
卷积层
加权&标准化
多头注意力层

图7

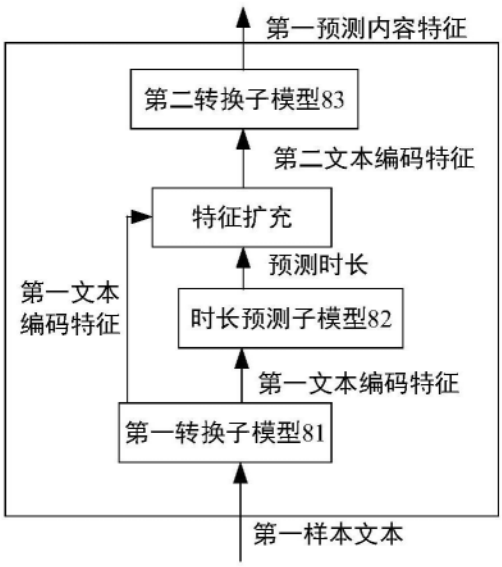

第一预测内容特征
第二转换子模型83
第二文本编码特征
特征扩充
预测时长
第一文本
编码特征
时长预测子模型82
第一文本编码特征
第一转换子模型81
第一样本文本

图8

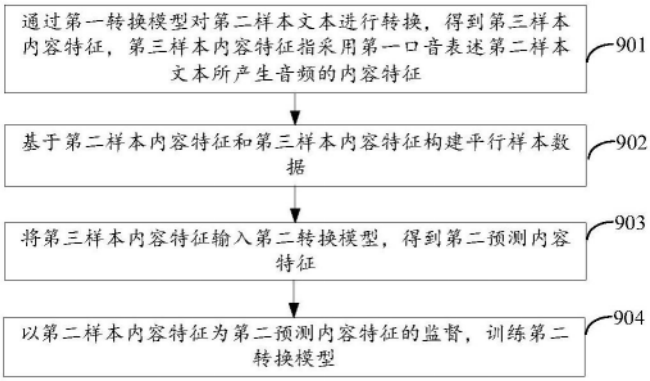

通过第一转换模型对第二样本文本进行转换，得到第三样本内容特征，第三样本内容特征指采用第一口音表述第二样本文本所产生音频的内容特征
901
基于第二样本内容特征和第三样本内容特征构建平行样本数据
902
将第三样本内容特征输入第二转换模型，得到第二预测内容特征
903
以第二样本内容特征为第二预测内容特征的监督，训练第二转换模型
904

图9

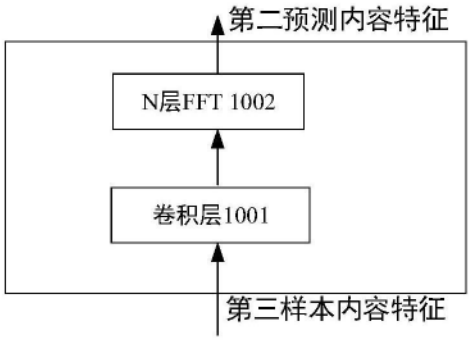

第二预测内容特征
N层FFT 1002
卷积层1001
第三样本内容特征

图10

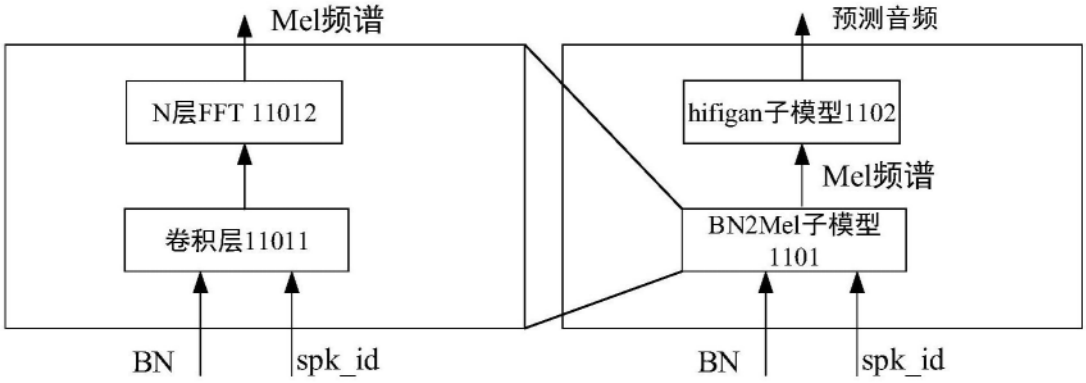


图11

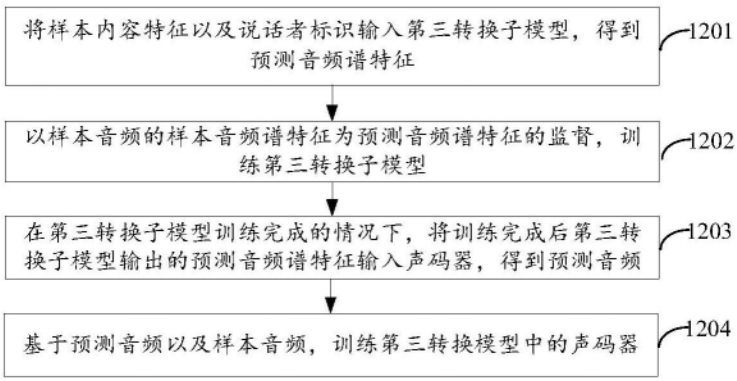


图12

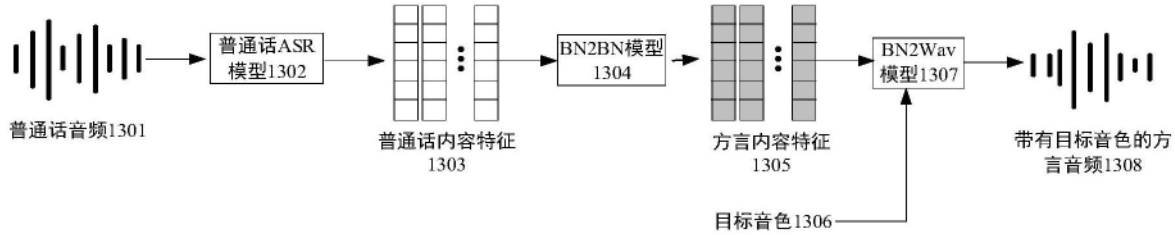


图13

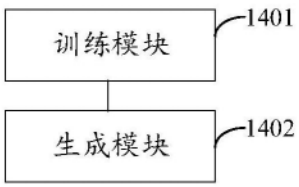


图14

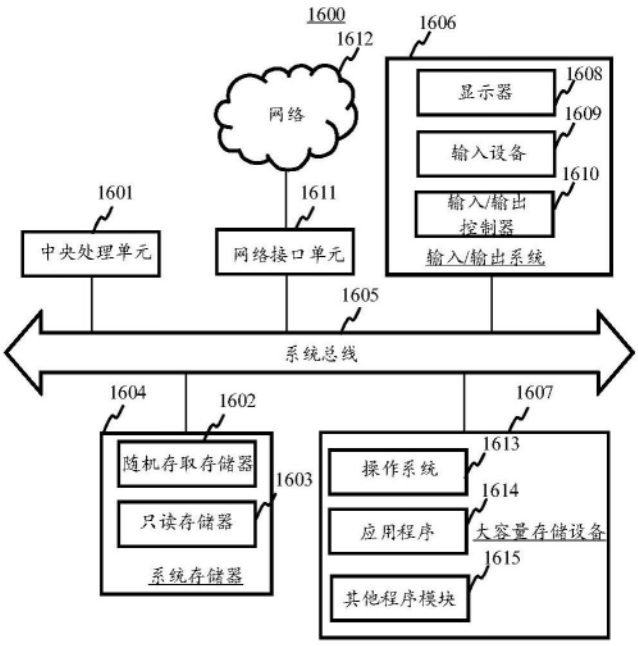

1600
1612
网络
1606
显示器
1608
输入设备
1609
输入/输出控制器
1610
输入/输出系统
1601
中央处理单元
1611
网络接口单元
1605
系统总线
1604
1602
随机存取存储器
只读存储器
1603
系统存储器
1607
操作系统
1613
应用程序
1614
大容量存储设备
其他程序模块
1615

图15