(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108550363 A
(43)申请公布日 2018.09.18

(21)申请号 201810565148.8

(22)申请日 2018.06.04

(71)申请人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号百度大厦

(72)发明人 顾宇 孙晓辉

(74)专利代理机构 北京鸿德海业知识产权代理事务所(普通合伙) 11412
代理人 袁媛

(51)Int.Cl.
G10L 13/08(2013.01)
G10L 13/047(2013.01)

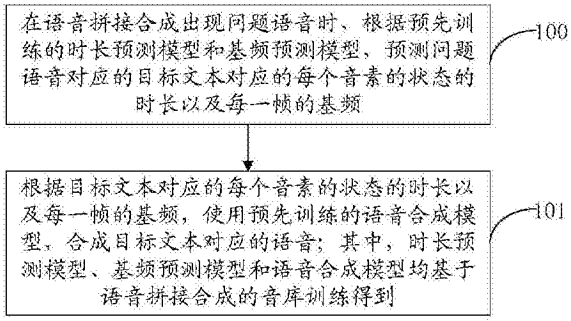
权利要求书2页 说明书10页 附图3页

(54)发明名称

语音合成方法及装置、计算机设备及可读介质

(57)摘要

本发明提供一种语音合成方法及装置、计算机设备及可读介质。其方法包括:在语音拼接合成出现问题语音时,根据预先训练的时长预测模型和基频预测模型,预测问题语音对应的目标文本对应的每个音素的状态的时长以及每一帧的基频;根据目标文本对应的每个音素的状态的时长以及每一帧的基频,使用预先训练的语音合成模型,合成目标文本对应的语音;时长预测模型、基频预测模型和语音合成模型均基于语音拼接合成的音库训练得到。本发明的技术方案,避免补充录制语料重新建库,可以有效地缩短问题语音修复的时间,节省问题语音修复成本;可以保证合成后的语音的自然度和连续性得到改善,且与拼接合成的语音音质相比,不会发生改变,不会影响用户的听感。



CN 108550363 A

1.一种语音合成方法，其特征在于，所述方法包括：

在语音拼接合成出现问题语音时，根据预先训练的时长预测模型和基频预测模型，预测所述问题语音对应的目标文本对应的每个音素的状态的时长以及每一帧的基频；

根据所述目标文本对应的所述每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成所述目标文本对应的语音；其中，所述时长预测模型、所述基频预测模型和所述语音合成模型均基于语音拼接合成的音库训练得到。

2.根据权利要求1所述的方法，其特征在于，根据预先训练的时长预测模型和基频预测模型，预测目标文本对应的每个音素的状态的时长以及每一帧的基频之前，所述方法还包括：

根据音库中的文本及对应的语音，训练所述时长预测模型、所述基频预测模型以及所述语音合成模型。

3.根据权利要求2所述的方法，其特征在于，根据音库中的文本及对应的语音，训练所述时长预测模型、所述基频预测模型以及所述语音合成模型，具体包括：

从所述音库中的文本及对应的语音中，提取数个训练文本及对应的训练语音；

分别从所述数个训练语音中提取各所述训练语音中每个音素对应的状态的时长以及每一帧对应的基频；

根据各所述训练文本及对应的所述训练语音中的每个音素对应的状态的时长，训练所述时长预测模型；

根据各所述训练文本及对应的所述训练语音中的每一帧对应的基频，训练所述基频预测模型；

根据各所述训练文本、对应的各所述训练语音、对应的各所述训练语音中的每个音素对应的状态的时长以及每一帧对应的基频，训练所述语音合成模型。

4.根据权利要求2所述的方法，其特征在于，根据预先训练的时长预测模型和基频预测模型，预测目标文本对应的每个音素的状态的时长以及每一帧的基频之前，所述方法还包括：

在使用所述音库进行语音拼接合成时，接收用户反馈的所述问题语音以及所述问题语音对应的所述目标文本。

5.根据权利要求2所述的方法，其特征在于，根据所述目标文本对应的所述每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成所述目标文本对应的语音之后，所述方法还包括：

将所述目标文本以及对应的合成的所述语音加入所述音库中。

6.根据权利要求1-5任一所述的方法，其特征在于，所述语音合成模型采用WaveNet模型。

7.一种语音合成装置，其特征在于，所述装置包括：

预测模块，用于在语音拼接合成出现问题语音时，根据预先训练的时长预测模型和基频预测模型，预测所述问题语音对应的目标文本对应的每个音素的状态的时长以及每一帧的基频；

合成模块，用于根据所述目标文本对应的所述每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成所述目标文本对应的语音；其中，所述时长预测模

型、所述基频预测模型和所述语音合成模型均基于语音拼接合成的音库训练得到。

8.根据权利要求7所述的装置，其特征在于，所述装置还包括：

训练模块，用于根据音库中的文本及对应的语音，训练所述时长预测模型、所述基频预测模型以及所述语音合成模型。

9.根据权利要求8所述的装置，其特征在于，所述训练模块，具体用于：

从所述音库中的文本及对应的语音中，提取数个训练文本及对应的训练语音；

分别从所述数个训练语音中提取各所述训练语音中每个音素对应的状态的时长以及每一帧对应的基频；

根据各所述训练文本及对应的所述训练语音中的每个音素对应的状态的时长，训练所述时长预测模型；

根据各所述训练文本及对应的所述训练语音中的每一帧对应的基频，训练所述基频预测模型；

根据各所述训练文本、对应的各所述训练语音、对应的各所述训练语音中的每个音素对应的状态的时长以及每一帧对应的基频，训练所述语音合成模型。

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：

接收模块，用于在使用所述音库进行语音拼接合成时，接收用户反馈的所述问题语音以及所述问题语音对应的所述目标文本。

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：

添加模块，用于将所述目标文本以及对应的合成的所述语音加入所述音库中。

12.根据权利要求7-11任一所述的装置，其特征在于，所述语音合成模型采用WaveNet模型。

13.一种计算机设备，其特征在于，所述设备包括：

一个或多个处理器；

存储器，用于存储一个或多个程序；

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-6中任一所述的方法。

14.一种计算机可读介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现如权利要求1-6中任一所述的方法。

语音合成方法及装置、计算机设备及可读介质

【技术领域】

[0001] 本发明涉及计算机应用技术领域，尤其涉及一种语音合成方法及装置、计算机设备及可读介质。

【背景技术】

[0002] 语音合成技术主要分为基于统计参数的技术和基于单元挑选的拼接合成技术两大类，这两大类语音合成方法都存在各自的优点，但也有各自相应的问题。

[0003] 例如，对于统计参数的语音合成技术，目前仅仅需要一个规模较小的音库，可以适用于离线场景下的语音合成任务，同时也可以应用于表现力合成、情感语音合成、说话人转换等任务中，这类方法合成的语音相对较为平稳，连续性较好，但是由于受到声学模型的建模能力有限以及统计平滑等效应的影响，统计参数合成的音质相对较差。不同于参数合成，拼接合成需要一个较大规模的音库，主要是应用于在线设备的语音合成任务中，拼接合成由于采用挑选音库中的波形片段，并通过特定算法拼接在一起，因此语音的音质较好，同自然语音较为接近，但是由于采用拼接的方式，很多不同语音单元之间的连续性较差。当给定合成的文本情况下，如果音库的候选单元的挑选不够精确或者特定词汇、短语未能被音库的语料所覆盖时，拼接合成语音会出现自然度和连续性较差的问题，会严重影响用户的听感。为了解决该技术问题，现有技术中采用补录音库的方式，再在音库中重新补充一些对应的语料，并重新建库以修复相应的问题。

[0004] 但是，现有技术中，从产品反馈的问题语音到再次邀请发音人补充录制语料，再进行重新建库，这是一个相对较长的迭代过程，问题语音的修复周期较长，无法达到即时修复的效果。

【发明内容】

[0005] 本发明提供了一种语音合成方法及装置、计算机设备及可读介质，用于快速修复拼接合成中自然度和连续性差的问题语音。

[0006] 本发明提供一种语音合成方法，所述方法包括：

[0007] 在语音拼接合成出现问题语音时，根据预先训练的时长预测模型和基频预测模型，预测所述问题语音对应的目标文本对应的每个音素的状态的时长以及每一帧的基频；

[0008] 根据所述目标文本对应的所述每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成所述目标文本对应的语音；其中，所述时长预测模型、所述基频预测模型和所述语音合成模型均基于语音拼接合成的音库训练得到。

[0009] 进一步可选地，如上所述的方法中，根据预先训练的时长预测模型和基频预测模型，预测目标文本对应的每个音素的状态的时长以及每一帧的基频之前，所述方法还包括：

[0010] 根据音库中的文本及对应的语音，训练所述时长预测模型、所述基频预测模型以及所述语音合成模型。

[0011] 进一步可选地，如上所述的方法中，根据音库中的文本及对应的语音，训练所述时

长预测模型、所述基频预测模型以及所述语音合成模型，具体包括：

[0012] 从所述音库中的文本及对应的语音中，提取数个训练文本及对应的训练语音；

[0013] 分别从所述数个训练语音中提取各所述训练语音中每个音素对应的状态的时长以及每一帧对应的基频；

[0014] 根据各所述训练文本及对应的所述训练语音中的每个音素对应的状态的时长，训练所述时长预测模型；

[0015] 根据各所述训练文本及对应的所述训练语音中的每一帧对应的基频，训练所述基频预测模型；

[0016] 根据各所述训练文本、对应的各所述训练语音、对应的各所述训练语音中的每个音素对应的状态的时长以及每一帧对应的基频，训练所述语音合成模型。

[0017] 进一步可选地，如上所述的方法中，根据预先训练的时长预测模型和基频预测模型，预测目标文本对应的每个音素的状态的时长以及每一帧的基频之前，所述方法还包括：

[0018] 在使用所述音库进行语音拼接合成时，接收用户反馈的所述问题语音以及所述问题语音对应的所述目标文本。

[0019] 进一步可选地，如上所述的方法中，根据所述目标文本对应的所述每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成所述目标文本对应的语音之后，所述方法还包括：

[0020] 将所述目标文本以及对应的合成的所述语音加入所述音库中。

[0021] 进一步可选地，如上所述的方法中，所述语音合成模型采用WaveNet模型。

[0022] 本发明提供一种语音合成装置，所述装置包括：

[0023] 预测模块，用于在语音拼接合成出现问题语音时，根据预先训练的时长预测模型和基频预测模型，预测所述问题语音对应的目标文本对应的每个音素的状态的时长以及每一帧的基频；

[0024] 合成模块，用于根据所述目标文本对应的所述每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成所述目标文本对应的语音；其中，所述时长预测模型、所述基频预测模型和所述语音合成模型均基于语音拼接合成的音库训练得到。

[0025] 进一步可选地，如上所述的装置中，还包括：

[0026] 训练模块，用于根据音库中的文本及对应的语音，训练所述时长预测模型、所述基频预测模型以及所述语音合成模型。

[0027] 进一步可选地，如上所述的装置中，所述训练模块，具体用于：

[0028] 从所述音库中的文本及对应的语音中，提取数个训练文本及对应的训练语音；

[0029] 分别从所述数个训练语音中提取各所述训练语音中每个音素对应的状态的时长以及每一帧对应的基频；

[0030] 根据各所述训练文本及对应的所述训练语音中的每个音素对应的状态的时长，训练所述时长预测模型；

[0031] 根据各所述训练文本及对应的所述训练语音中的每一帧对应的基频，训练所述基频预测模型；

[0032] 根据各所述训练文本、对应的各所述训练语音、对应的各所述训练语音中的每个音素对应的状态的时长以及每一帧对应的基频，训练所述语音合成模型。

[0033] 进一步可选地，如上所述的装置中，还包括：

[0034] 接收模块，用于在使用所述音库进行语音拼接合成时，接收用户反馈的所述问题语音以及所述问题语音对应的所述目标文本。

[0035] 进一步可选地，如上所述的装置中，还包括：

[0036] 添加模块，用于将所述目标文本以及对应的合成的所述语音加入所述音库中。

[0037] 进一步可选地，如上所述的装置中，所述语音合成模型采用WaveNet模型。

[0038] 本发明还提供一种计算机设备，所述设备包括：

[0039] 一个或多个处理器；

[0040] 存储器，用于存储一个或多个程序；

[0041] 当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如上所述的语音合成方法。

[0042] 本发明还提供一种计算机可读介质，其上存储有计算机程序，该程序被处理器执行时实现如上所述的语音合成方法。

[0043] 本发明的语音合成方法及装置、计算机设备及可读介质，通过在语音拼接合成出现问题语音时，根据预先训练的时长预测模型和基频预测模型，预测问题语音对应的目标文本对应的每个音素的状态的时长以及每一帧的基频；根据目标文本对应的每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成目标文本对应的语音；其中，时长预测模型、基频预测模型和语音合成模型均基于语音拼接合成的音库训练得到。本发明的技术方案，可以在语音拼接合成出现问题语音时，基于上述方式来实现问题语音的修复，避免补充录制语料重新建库，可以有效地缩短问题语音修复的时间，节省问题语音修复成本，提高问题语音修复的效率；而且本发明的技术方案中，由于时长预测模型、基频预测模型和语音合成模型均基于语音拼接合成的音库训练得到，可以保证模型合成后的语音的自然度和连续性差，且与拼接合成的语音音质相比，不会发生改变，不会影响用户的听感。

【附图说明】

[0044] 图1为本发明的语音合成方法实施例一的流程图。

[0045] 图2为本发明的语音合成方法实施例二的流程图。

[0046] 图3为本发明的语音合成装置实施例一的结构图。

[0047] 图4为本发明的语音合成装置实施例二的结构图。

[0048] 图5为本发明的计算机设备实施例的结构图。

[0049] 图6为本发明提供的一种计算机设备的示例图。

【具体实施方式】

[0050] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

[0051] 图1为本发明的语音合成方法实施例一的流程图。如图1所示，本实施例的语音合成方法，具体可以包括如下步骤：

[0052] 100、在语音拼接合成出现问题语音时，根据预先训练的时长预测模型和基频预测

模型，预测问题语音对应的目标文本对应的每个音素的状态的时长以及每一帧的基频；

[0053] 101、根据目标文本对应的每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成目标文本对应的语音；其中，时长预测模型、基频预测模型和语音合成模型均基于语音拼接合成的音库训练得到。

[0054] 本实施例的语音合成方法的执行主体为语音合成装置。具体地，在语音拼接合成的过程中，若待合成文本未能被音库的语料完全覆盖时，会导致拼接合成语音出现自然度和连续性较差的问题，而现有技术中修复该问题的需要补充录制语料并重新建库，导致问题语音的修复周期较长。为了解决该问题，本实施例中，采用语音合成装置实现对这部分待合成文本的语音合成，作为现有的语音拼接合成过程中出现问题语音时的补充方案，实现从另一个角度实现语音合成，以有效地缩短问题语音的修复周期。

[0055] 具体地，本实施例的语音合成方法中，需要预先训练的时长预测模型和基频预测模型。其中该时长预测模型用于预测目标文本中的每一个音素的状态的时长。其中音素为语音中的一个最小单元，例如在中文的读音中，一个声母或者韵母可以分别作为一个音素。在其他语言的读音中，每一个发音也相当于一个音素。本实施例中，每一个音素可以按照隐马尔可夫模型切分为5个状态，状态时长为停留在状态的时间长度。本实施例中预先训练的时长预测模型可以预测目标文本中的每个音素的所有状态时长。另外，本实施例中，还预先训练有基频预测模型，该基频预测模型可以预测目标文本的读音中每一帧的基频。

[0056] 本实施例的目标文本对应的每个音素的状态时长以及每一帧的基频为语音合成的必要特征。具体地，可以将目标文本对应的每个音素的状态时长以及每一帧的基频输入至预先训练的语音合成模型中，该语音合成模型便可以合成并输出该目标文本对应的语音。这样，当拼接合成时出现问题自然度和连续性差的问题时，可以直接使用本实施例的方案进行语音合成。由于本实施例的语音合成方案中的时长预测模型、基频预测模型和语音合成模型均基于语音拼接合成的音库训练得到，所以，可以保证合成的语音的音质与语音拼接合成的音库中的音质相同，即使得合成的读音与拼接的读音听起来同一个发音人的语音，从而可以保证用户的听感，增强用户的使用体验度。而且本实施例的语音合成方案中的时长预测模型、基频预测模型和语音合成模型都是预先训练的，所以在修复问题语音时，可以到达即时修复的效果。

[0057] 本实施例的语音合成方法，通过根据预先训练的时长预测模型和基频预测模型，预测目标文本对应的每个音素的状态的时长以及每一帧的基频；根据目标文本对应的每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成目标文本对应的语音；其中，时长预测模型、基频预测模型和语音合成模型均基于语音拼接合成的音库训练得到。本实施例的技术方案，可以在语音拼接合成出现问题语音时，基于上述方式来实现问题语音的修复，避免补充录制语料重新建库，可以有效地缩短问题语音修复的时间，节省问题语音修复成本，提高问题语音修复的效率；而且本实施例的技术方案中，由于时长预测模型、基频预测模型和语音合成模型均基于语音拼接合成的音库训练得到，可以保证模型合成后的语音的自然度和连续性，且与拼接合成的语音音质相比，不会发生改变，不会影响用户的听感。

[0058] 图2为本发明的语音合成方法实施例二的流程图。如图2所示，本实施例的语音合成方法，在上述图1所示实施例的技术方案的基础上，进一步更加详细地介绍本发明的技术

方案。如图2所示，本实施例的语音合成方法，具体可以包括如下步骤：

[0059] 200、根据音库中的文本及对应的语音，训练时长预测模型、基频预测模型以及语音合成模型；

[0060] 具体地，该步骤200具体可以包括如下步骤：

[0061] (a)从音库中的文本及对应的语音中，提取数个训练文本及对应的训练语音；

[0062] (b)分别从数个训练语音中提取各训练语音中每个音素对应的状态的时长以及每一帧对应的基频；

[0063] (c)根据各训练文本及对应的训练语音中的每个音素对应的状态的时长，训练时长预测模型；

[0064] (d)根据各训练文本及对应的训练语音中的每一帧对应的基频，训练基频预测模型；

[0065] (e)根据各训练文本、对应的各训练语音、对应的各训练语音中的每个音素对应的状态的时长以及每一帧对应的基频，训练语音合成模型。

[0066] 本实施例的语音拼接合成所使用的音库可以包括足量的原始语料，该原始语料中可以包括原始文本以及对应的原始语音，例如可以包括20个小时的原始语音。首先，从音库中提取数个训练文本及对应的训练语音，例如每个训练文本可以为一句话。然后分别从数个训练语音中按照隐马尔可夫模型提取各训练语音中每个音素对应的状态的时长，同时还可以提取数个训练语音中每个训练语音中的每一帧对应的基频。然后分别训练三个模型。本实施例的数个训练文本及对应的训练语音的具体数量可以根据实际需求来设置，例如可以为上万数量的训练文本以及对应的训练语音。

[0067] 例如根据各训练文本及对应的训练语音中的每个音素对应的状态的时长，训练时长预测模型。训练之前，可以为该时长预测模型设置初始参数。然后输入训练文本，由时长预测模型预测该训练文本对应的训练语音中的每个音素对应的状态的预测时长；接下类将时长预测模型预测的该训练文本对应的训练语音中的每个音素对应的状态的预测时长，与对应的训练语音中的每个音素对应的状态的真实时长进行比对，判断两者的差值是否在一个预设范围内，若不在，则调整时长预测模型的参数，使得两者的差值落入一个预设范围内。采用多条训练文本及对应的训练语音中的每个音素对应的状态的时长，不断地对时长预测模型进行训练，确定时长预测模型的参数，从而确定时长预测模型，时长预测模型训练完毕。

[0068] 另外，具体可以根据各训练文本及对应的训练语音中的每一帧对应的基频，训练基频预测模型。同理，训练之前，可以为该基频预测模型设置初始参数。由基频预测模型预测该训练文本对应的训练语音中的每一帧对应的预测基频；接下类判断基频预测模型预测的每一帧的基频，与对应的训练语音中的每一帧的真实基频进行比对，判断两者的差值是否在一个预设范围内，若不在，则调整基频预测模型的参数，使得两者的差值落入一个预设范围内。采用多条训练文本及对应的训练语音中每一帧对应的基频，不断地对基频预测模型进行训练，确定基频预测模型的参数，从而确定基频预测模型，基频预测模型训练完毕。

[0069] 而且，还可以根据各训练文本、对应的各训练语音、对应的各训练语音中的每个音素对应的状态的时长以及每一帧对应的基频，训练语音合成模型。本实施例的语音合成模型可以采用WaveNet模型。该WaveNet模型为DeepMind团队2016年提出的具有波形建模能力

的模型，该WaveNet模型自提出以来，受到工业界和学术界的广泛关注。

[0070] 在该语音合成模型如WaveNet模型中，将每条训练文本的训练语音中的每个音素对应的状态的时长以及每一帧对应的基频作为合成语音的必要特征。训练之前，为该WaveNet模型设置初始参数。训练时，将各训练文本、对应的各训练语音中的每个音素对应的状态的时长以及每一帧对应的基频输入至该WaveNet模型中，WaveNet模型根据输入的特征输出合成后的语音；然后计算该合成后的语音与训练语音的交叉熵；接着采用梯度下降方法调整WaveNet模型的参数，使得该交叉熵达到一个极小值，即表示WaveNet模型合成的语音与对应的训练语音足够接近。按照上述方式，采用多条训练文本、对应的多条训练语音、以及对应的各训练语音中的每个音素对应的状态的时长和每一帧对应的基频，不断地对WaveNet模型进行训练，确定WaveNet模型的参数，从而确定WaveNet模型，WaveNet模型的训练完毕。

[0071] 本实施例的上述训练时长预测模型、基频预测模型以及语音合成模型可以为离线训练的过程，以得到上述三个模型，以在拼接语音合成时出现问题时，进行在线使用。

[0072] 201、在使用音库进行语音拼接合成时，判断是否接收到用户反馈的问题语音以及问题语音对应的目标文本；若是，执行步骤202；否则继续使用音库进行语音拼接合成。

[0073] 202、确定采用语音拼接技术根据音库拼接的目标文本的语音为问题语音；执行步骤203；

[0074] 由于在语音拼接合成时，若音库中缺少目标文本的语料时，会造成拼接的语音连续性和自然性较差，此时合成的语音为问题语音，通常造成用户的无法正常使用。

[0075] 203、根据预先训练的时长预测模型和基频预测模型，预测目标文本对应的每个音素的状态的时长以及每一帧的基频；执行步骤204；

[0076] 204、根据目标文本对应的每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成目标文本对应的语音；执行步骤205；

[0077] 步骤203和步骤204可以参考上述图1所示实施例的步骤100和步骤101，在此不再赘述。

[0078] 205、将目标文本以及对应的合成的语音加入音库中，以对音库进行升级。

[0079] 经过上述处理，可以将合成该目标文本对应的语音，然后可以将该语音加入至音库中，这样，后续再使用音库进行语音拼接合成时，可以提高语音拼接合成的自然性和连续性。而只有在出现问题语音的时候，再采用本实施例的方式合成语音，且合成的语音与音库中的原始语音具有相同的音质，使得用户听起来时同一发音人发出的，不会影响用户的听感。而且通过本实施例的方式，可以不断地扩充音库中的语料，使得后续使用语音拼接合成的效率更高；且本实施例的技术方案，通过更新音库，不仅使得音库得以升级，还可以使得使用更新后的音库的语音拼接合成系统的服务进行了升级，能够满足更多的语音拼接合成需求。

[0080] 本实施例的语音合成方法，可以在语音拼接合成出现问题语音时，基于上述方式来实现问题语音的修复，避免补充录制语料重新建库，可以有效地缩短问题语音修复的时间，节省问题语音修复成本，提高问题语音修复的效率；而且本实施例的技术方案中，由于时长预测模型、基频预测模型和语音合成模型均基于语音拼接合成的音库训练得到，可以保证模型合成后的语音的自然度和连续性差，且与拼接合成的语音音质相比，不会发生改

变，不会影响用户的听感。

[0081] 图3为本发明的语音合成装置实施例一的结构图。如图4所示，本实施例的语音合成装置，具体可以包括：

[0082] 预测模块10用于在语音拼接合成出现问题语音时，根据预先训练的时长预测模型和基频预测模型，预测问题语音对应的目标文本对应的每个音素的状态的时长以及每一帧的基频；

[0083] 合成模块11用于根据预测模块10预测的目标文本对应的每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成目标文本对应的语音；其中，时长预测模型、基频预测模型和语音合成模型均基于语音拼接合成的音库训练得到。

[0084] 本实施例的语音合成装置，通过采用上述模块实现语音合成的实现原理以及技术效果与上述相关方法实施例的实现相同，详细可以参考上述相关方法实施例的记载，在此不再赘述。

[0085] 图4为本发明的语音合成装置实施例二的结构图。如图4所示，本实施例的语音合成装置，在上述图3所示实施例的技术方案的基础上，具体可以包括：

[0086] 如图4所示，本实施例的语音合成装置，还包括：训练模块12用于根据音库中的文本及对应的语音，训练时长预测模型、基频预测模型以及语音合成模型。

[0087] 对应地，预测模块10用于根据训练模块12预先训练的时长预测模型和基频预测模型，预测问题语音对应的目标文本对应的每个音素的状态的时长以及每一帧的基频；

[0088] 对应地，合成模块11用于根据预测模块10预测的目标文本对应的每个音素的状态的时长以及每一帧的基频，使用训练模块12预先训练的语音合成模型，合成目标文本对应的语音；

[0089] 进一步可选地，如图4所示，本实施例的语音合成装置，训练模块12具体用于：

[0090] 从音库中的文本及对应的语音中，提取数个训练文本及对应的训练语音；

[0091] 分别从数个训练语音中提取各训练语音中每个音素对应的状态的时长以及每一帧对应的基频；

[0092] 根据各训练文本及对应的训练语音中的每个音素对应的状态的时长，训练时长预测模型；

[0093] 根据各训练文本及对应的训练语音中的每一帧对应的基频，训练基频预测模型；

[0094] 根据各训练文本、对应的各训练语音、对应的各训练语音中的每个音素对应的状态的时长以及每一帧对应的基频，训练语音合成模型。

[0095] 进一步可选地，如图4所示，本实施例的语音合成装置，还包括：

[0096] 接收模块13用于在使用音库进行语音拼接合成时，接收用户反馈的问题语音以及问题语音对应的所述目标文本。

[0097] 对应地，接收模块13可以触发预测模块10，在接收模块13接收到用户反馈的问题语音后，触发预测模块10根据预先训练的时长预测模型和基频预测模型，预测目标文本对应的每个音素的状态的时长以及每一帧的基频。

[0098] 进一步可选地，如图4所示，本实施例的语音合成装置，还包括：

[0099] 添加模块14用于将目标文本以及合成模块11合成的对应的语音加入音库中。

[0100] 进一步可选地，本实施例的语音合成装置中，语音合成模型采用WaveNet模型。

[0101] 本实施例的语音合成装置，通过采用上述模块实现语音合成的实现原理以及技术效果与上述相关方法实施例的实现相同，详细可以参考上述相关方法实施例的记载，在此不再赘述。

[0102] 图5为本发明的计算机设备实施例的结构图。如图5所示，本实施例的计算机设备，包括：一个或多个处理器30，以及存储器40，存储器40用于存储一个或多个程序，当存储器40中存储的一个或多个程序被一个或多个处理器30执行，使得一个或多个处理器30实现如上图1-图2所示实施例的语音合成方法。图5所示实施例中以包括多个处理器30为例。

[0103] 例如，图6为本发明提供的一种计算机设备的示例图。图6示出了适于用来实现本发明实施方式的示例性计算机设备12a的框图。图6显示的计算机设备12a仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0104] 如图6所示，计算机设备12a以通用计算设备的形式表现。计算机设备12a的组件可以包括但不限于：一个或者多个处理器16a，系统存储器28a，连接不同系统组件(包括系统存储器28a和处理器16a)的总线18a。

[0105] 总线18a表示几类总线结构中的一种或多种，包括存储器总线或者存储器控制器，外围总线，图形加速端口，处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说，这些体系结构包括但不限于工业标准体系结构(ISA)总线，微通道体系结构(MAC)总线，增强型ISA总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0106] 计算机设备12a典型地包括多种计算机系统可读介质。这些介质可以是任何能够被计算机设备12a访问的可用介质，包括易失性和非易失性介质，可移动的和不可移动的介质。

[0107] 系统存储器28a可以包括易失性存储器形式的计算机系统可读介质，例如随机存取存储器(RAM)30a和/或高速缓存存储器32a。计算机设备12a可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例，存储系统34a可以用于读写不可移动的、非易失性磁介质(图6未显示，通常称为“硬盘驱动器”)。尽管图6中未示出，可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器，以及对可移动非易失性光盘(例如CD-ROM，DVD-ROM或者其它光介质)读写的光盘驱动器。在这些情况下，每个驱动器可以通过一个或者多个数据介质接口与总线18a相连。系统存储器28a可以包括至少一个程序产品，该程序产品具有一组(例如至少一个)程序模块，这些程序模块被配置以执行本发明上述图1-图4各实施例的功能。

[0108] 具有一组(至少一个)程序模块42a的程序/实用工具40a，可以存储在例如系统存储器28a中，这样的程序模块42a包括——但不限于——操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块42a通常执行本发明所描述的上述图1-图4各实施例中的功能和/或方法。

[0109] 计算机设备12a也可以与一个或多个外部设备14a(例如键盘、指向设备、显示器24a等)通信，还可与一个或者多个使得用户能与该计算机设备12a交互的设备通信，和/或与使得该计算机设备12a能与一个或多个其它计算设备进行通信的任何设备(例如网卡，调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口22a进行。并且，计算机设备12a还可以通过网络适配器20a与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。如图所示，网络适配器20a通过总线18a与计算机设备12a的其

它模块通信。应当明白，尽管图中未示出，可以结合计算机设备12a使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理器、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0110] 处理器16a通过运行存储在系统存储器28a中的程序，从而执行各种功能应用以及数据处理，例如实现上述实施例所示的语音合成方法方法。

[0111] 本发明还提供一种计算机可读介质，其上存储有计算机程序，该程序被处理器执行时实现如上述实施例所示的语音合成方法方法。

[0112] 本实施例的计算机可读介质可以包括上述图6所示实施例中的系统存储器28a中的RAM30a、和/或高速缓存存储器32a、和/或存储系统34a。

[0113] 随着科技的发展，计算机程序的传播途径不再受限于有形介质，还可以直接从网络下载，或者采用其他方式获取。因此，本实施例中的计算机可读介质不仅可以包括有形的介质，还可以包括无形的介质。

[0114] 本实施例的计算机可读介质可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0115] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括——但不限于——电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0116] 计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括——但不限于——无线、电线、光缆、RF等等，或者上述的任意合适的组合。

[0117] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如"C"语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0118] 在本发明所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

[0119] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0120] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

[0121] 上述以软件功能单元的形式实现的集成的单元，可以存储在一个计算机可读取存储介质中。上述软件功能单元存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)或处理器(processor)执行本发明各个实施例所述方法的部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0122] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

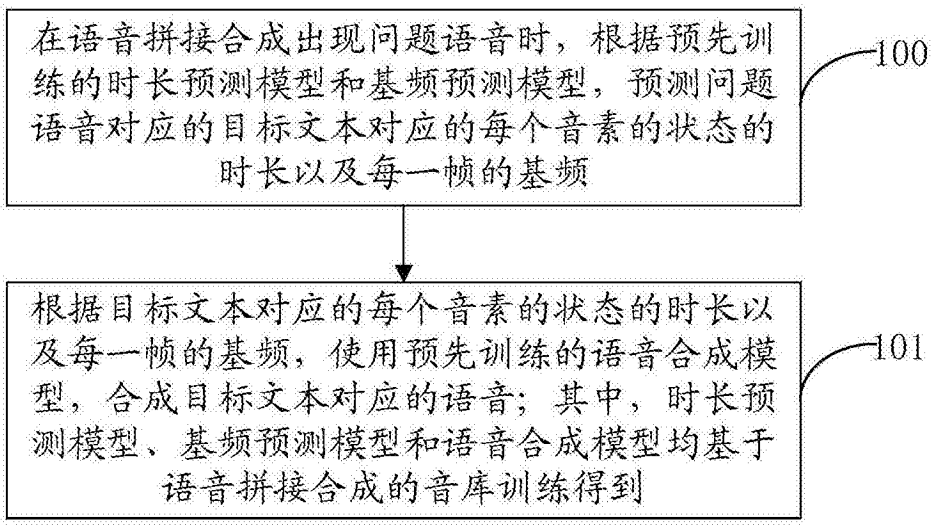

在语音拼接合成出现问题语音时，根据预先训练的时长预测模型和基频预测模型，预测问题语音对应的目标文本对应的每个音素的状态的时长以及每一帧的基频
100
根据目标文本对应的每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成目标文本对应的语音；其中，时长预测模型、基频预测模型和语音合成模型均基于语音拼接合成的音库训练得到
101

图1

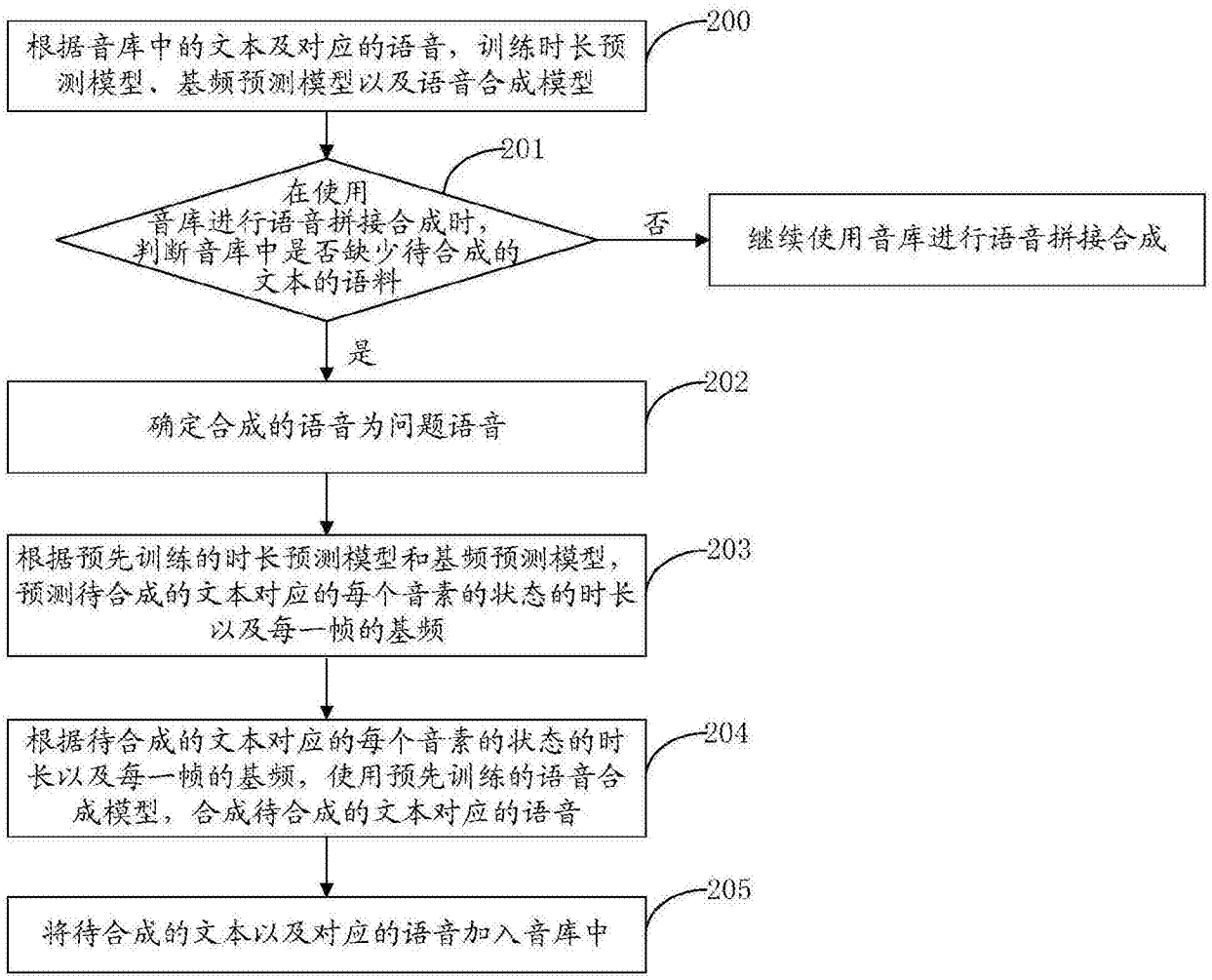

根据音库中的文本及对应的语音，训练时长预测模型、基频预测模型以及语音合成模型
200
在使用音库进行语音拼接合成时，判断音库中是否缺少待合成的文本的语料
201
否
继续使用音库进行语音拼接合成
是
确定合成的语音为问题语音
202
根据预先训练的时长预测模型和基频预测模型，预测待合成的文本对应的每个音素的状态的时长以及每一帧的基频
203
根据待合成的文本对应的每个音素的状态的时长以及每一帧的基频，使用预先训练的语音合成模型，合成待合成的文本对应的语音
204
将待合成的文本以及对应的语音加入音库中
205

图2

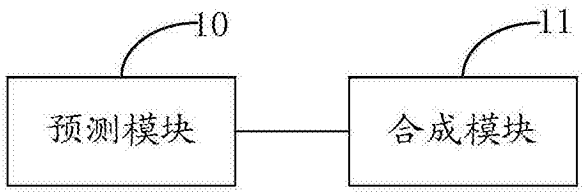

10
11
预测模块
合成模块

图3

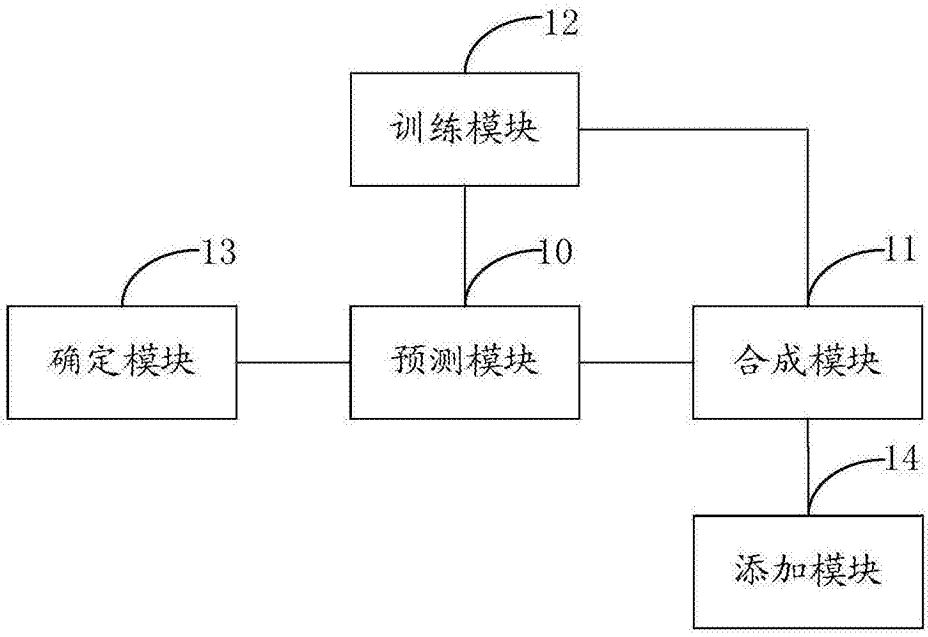

12
训练模块
13
10
11
确定模块
预测模块
合成模块
14
添加模块

图4

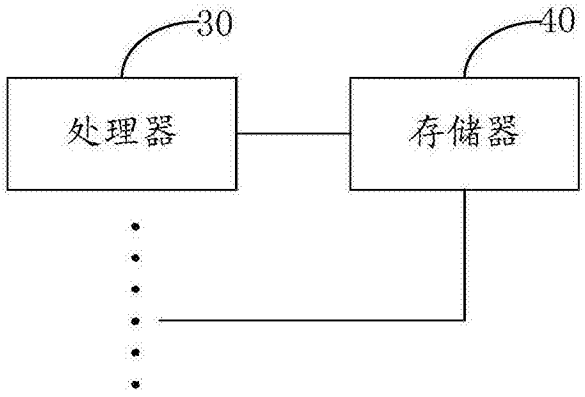

30
40
处理器
存储器

图5

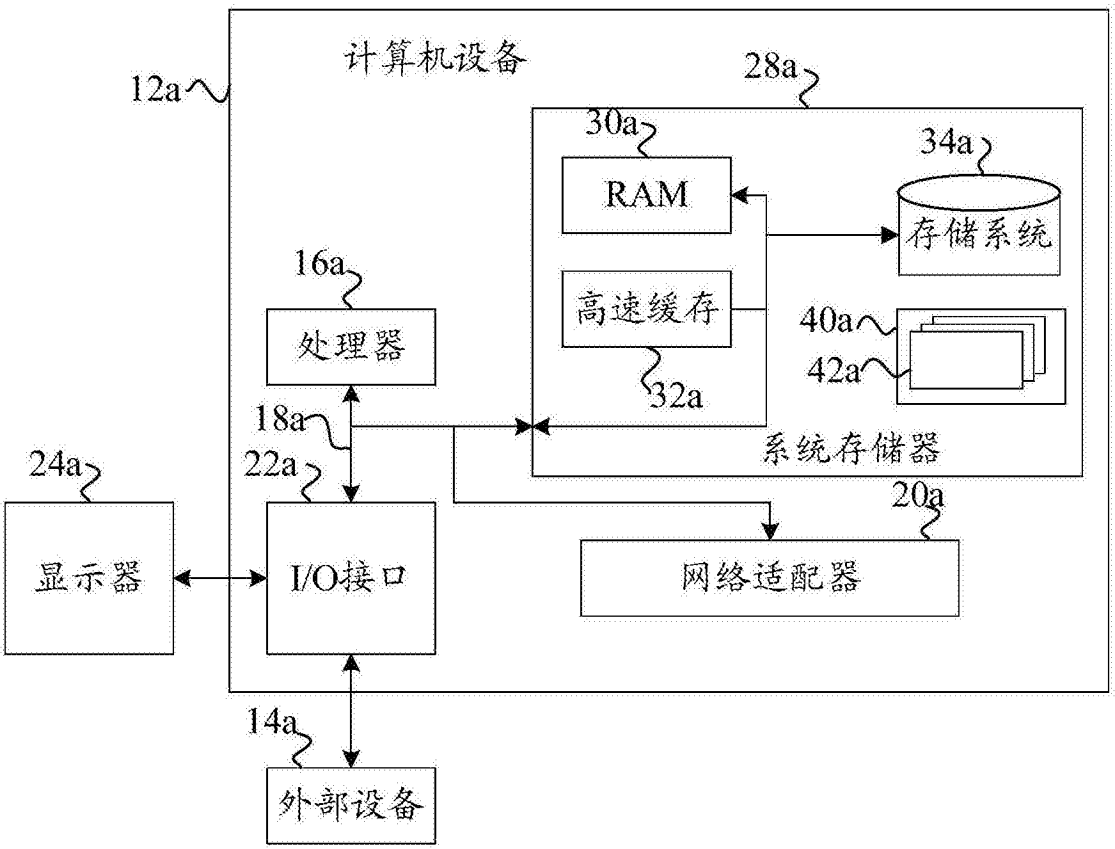

计算机设备
12a
28a
30a
RAM
34a
存储系统
16a
处理器
高速缓存
40a
42a
32a
18a
系统存储器
24a
22a
20a
显示器
I/O接口
网络适配器
14a
外部设备

图6