(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112837669 B
(45)授权公告日 2023.10.24

(21)申请号 202010437019.8
(22)申请日 2020.05.21
(65)同一申请的已公布的文献号
申请公布号 CN 112837669 A
(43)申请公布日 2021.05.25
(73)专利权人 腾讯科技(深圳)有限公司
地址 518064 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 林诗伦 蒙力 苏文超 唐宗尧 李新辉 卢鲤
(74)专利代理机构 深圳市智圈知识产权代理事务所(普通合伙) 44351
专利代理师 韩绍君
(51)Int.Cl.
G10L 13/02(2013.01)
G10L 13/04(2013.01)

(56)对比文件
CN 105355194 A,2016.02.24
US 2020066253 A1,2020.02.27
CN 105489216 A,2016.04.13
CN 110288973 A,2019.09.27
CN 110808027 A,2020.02.18
CN 106156857 A,2016.11.23

审查员 林登樟

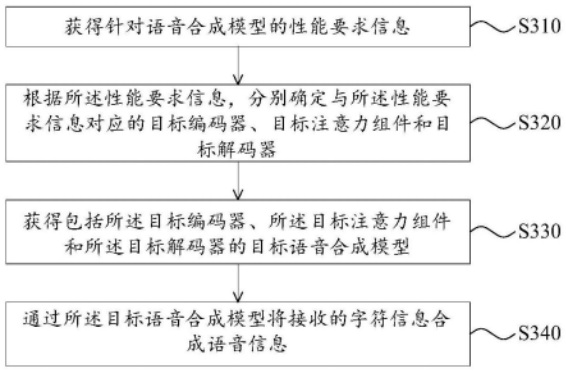
权利要求书3页 说明书19页 附图10页

(54)发明名称

语音合成方法、装置及服务器

(57)摘要

本申请公开了一种语音合成方法、装置及服务器，涉及人工智能技术领域。其中，该方法包括：获得针对语音合成模型的性能要求信息；根据性能要求信息，分别确定与性能要求信息对应的目标编码器、目标注意力组件和目标解码器；获得包括目标编码器、目标注意力组件和目标解码器的目标语音合成模型；通过目标语音合成模型将接收的字符信息合成为语音信息。如此，可以根据性能要求的不同来获得对应的语音合成模型，从而基于该语音合成模型提供符合该性能要求的语音合成服务。



CN 112837669 B

1.一种语音合成方法,其特征在于,所述方法包括:

获得针对语音合成模型的性能要求信息;

根据所述性能要求信息,分别确定与所述性能要求信息对应的目标编码器、目标注意力组件和目标解码器;若所述性能要求信息包括第一性能指标信息,确定与所述性能要求信息对应的目标编码器的步骤包括:将具有残差网络结构的编码器确定为目标编码器,其中,所述残差网络结构包括依次连接的第一编码层和第二编码层,所述第一编码层的输出信息被叠加至所述第二编码层的输出信息,所述第一性能指标信息为表示对不同字符信息的表征能力良好的信息或者表示对不同字符信息的区分能力好的信息;

获得包括所述目标编码器、所述目标注意力组件和所述目标解码器的目标语音合成模型;

通过所述目标语音合成模型将接收的字符信息合成为语音信息。

2.根据权利要求1所述的方法,其特征在于,所述获得包括所述目标编码器、所述目标注意力组件和所述目标解码器的目标语音合成模型,包括:

确定语音合成模型的目标框架;

按照所述目标框架对所述目标编码器、所述目标注意力组件和所述目标解码器进行组合,得到所述目标语音合成模型。

3.根据权利要求1所述的方法,其特征在于,所述目标编码器通过以下方式对所述字符信息进行处理:

通过所述第一编码层,按照接收次序对所述字符信息中的每个字符进行编码,得到第一字符特征向量;

通过所述第二编码层处理所述第一字符特征向量,得到第二字符特征向量;

拼接所述第一字符特征向量和所述第二字符特征向量,得到当前编码次序的字符对应的目标字符特征向量。

4.根据权利要求1所述的方法,其特征在于,所述根据所述性能要求信息,分别确定与所述性能要求信息对应的目标编码器、目标注意力组件和目标解码器,包括:

若所述性能要求信息包括第二性能指标信息,将包括参数拟合层的注意力组件确定为所述目标注意力组件,其中,所述参数拟合层用于拟合高斯混合模型的参数特征信息,所述高斯混合模型是所述目标编码器在每个编码次序编码的字符与所述目标解码器在当前解码次序解码的字符的相关程度所服从的概率分布模型,所述第二性能指标信息为表示能够稳定识别超长字符信息的信息,所述超长字符信息为字符数量达到第二数量的字符信息。

5.根据权利要求4所述的方法,其特征在于,所述目标注意力组件通过以下方式确定所述高斯混合模型的模型参数:

通过所述参数拟合层处理所述目标编码器在各编码次序的隐藏状态以及所述目标解码器在当前解码次序的隐藏状态,得到所述高斯混合模型的第一参数特征信息、第二参数特征信息和第三参数特征信息;

通过软最大化函数处理所述第一参数特征信息,得到所述高斯混合模型中每个高斯分布的权重;

根据所述第二参数特征信息得到所述高斯混合模型中每个高斯分布的方差;

通过软加函数处理所述第三参数特征信息,得到所述高斯混合模型中每个高斯分布的

均值。

6.根据权利要求5所述的方法，其特征在于，所述目标注意力组件通过以下方式处理所述目标编码器输出的目标字符特征向量：

根据所述高斯混合模型中每个高斯分布的权重、方差和均值，得到所述高斯混合模型在当前解码次序的概率分布函数；

根据所述概率分布函数和所述目标编码器在每个编码次序输出的目标字符特征向量，得到该目标字符特征向量在当前解码次序的注意力得分；

根据所述目标编码器在各编码次序输出的目标字符向量及每个目标字符向量在当前解码次序的注意力得分，得到所述目标解码器在当前解码次序的注意力向量。

7.根据权利要求1所述的方法，其特征在于，所述根据所述性能要求信息，分别确定结构与所述性能要求信息对应的目标编码器、目标注意力组件和目标解码器，包括：

若所述性能要求信息包括第三性能指标信息，将包括依次连接的循环门单元GRU层和长短时记忆网络LSTM层的解码器确定为所述目标解码器，所述第三性能指标信息为表示算力成本低的信息。

8.根据权利要求7所述的方法，其特征在于，所述目标解码器通过以下方式获得预测声学特征信息：

获取当前解码次序的注意力向量及所述当前解码次序的前一解码次序的注意力向量；

通过所述GRU层，对所述前一解码次序的注意力向量、所述当前解码次序的目标字符特征向量及所述目标解码器在前一解码次序的解码信息进行处理，得到第一声学特征向量；

通过所述LSTM层处理所述第一声学特征及所述当前解码次序的注意力向量，得到第二声学特征向量；

根据所述第二声学特征向量得到并输出所述目标解码器在所述当前解码次序的预测声学特征信息。

9.根据权利要求1或2所述的方法，其特征在于，在所述通过所述目标语音合成模型将接收的字符信息合成为语音信息之前，所述方法还包括：

获得针对语音合成模型的音色要求信息；

根据所述音色要求信息获取声音数据；

基于所述声音数据对所述目标语音合成模型进行模型训练，使所述目标语音合成模型的第一损失函数达到优化条件。

10.根据权利要求9所述的方法，其特征在于，所述第一损失函数通过如下方式建立：

获取所述目标编码器在各编码次序输出的目标字符向量各自在当前解码次序的注意力得分，得到一注意力得分序列；

确定所述注意力得分序列的熵；

将所述熵叠加至第二损失函数，得到所述第一损失函数。

11.根据权利要求10所述的方法，其特征在于，所述将所述熵叠加至第二损失函数，包括：

将所述熵与目标权重的乘积叠加至所述第二损失函数，其中，所述目标权重在所述模型训练的过程中随迭代次数的增大而增大。

12.一种语音合成装置，其特征在于，所述装置包括：

信息获得模块，用于获得针对语音合成模型的性能要求信息；

确定模块，用于根据所述性能要求信息，分别确定与所述性能要求信息对应的目标编码器、目标注意力组件和目标解码器，若所述性能要求信息包括第一性能指标信息，确定模块还用于将具有残差网络结构的编码器确定为目标编码器，其中，所述残差网络结构包括依次连接的第一编码层和第二编码层，所述第一编码层的输出信息被叠加至所述第二编码层的输出信息，所述第一性能指标信息为表示对不同字符信息的表征能力良好的信息或者表示对不同字符信息的区分能力好的信息；

模型获得模块，用于获得包括所述目标编码器、所述目标注意力组件和所述目标解码器的目标语音合成模型；

语音合成模块，用于通过所述目标语音合成模型将接收的字符信息合成为语音信息。

13.一种服务器，其特征在于，包括：

一个或多个处理器；

存储器；

一个或多个程序，其中所述一个或多个程序被存储在所述存储器中并被配置为由所述一个或多个处理器执行，所述一个或多个程序配置用于执行如权利要求1-11中任意一项所述的方法。

14.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有程序代码，所述程序代码可被处理器调用执行如权利要求1-11中任意一项所述的方法。

语音合成方法、装置及服务器

技术领域

[0001] 本申请涉及人工智能技术领域，更具体地，涉及一种语音合成方法、装置及服务器。

背景技术

[0002] 人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术及机器学习/深度学习等几大方向。

[0003] 其中，语音处理技术的一个重要分支是语音合成技术(Text to Speech,TTS)，即用于将文字信息转换成语音信息的技术。目前，深度学习在语音合成技术领域得到了广泛应用，其中应用比较广泛的是基于深度学习的端到端语音合成系统。然而，目前的基于深度学习的端到端语音合成系统是一个通用型结构，难以适应不同的需求。

发明内容

[0004] 本申请提出了一种语音合成方法、装置及服务器，可以改善上述问题。

[0005] 一方面，本申请实施例提供了一种语音合成方法，该方法包括：获得针对语音合成模型的性能要求信息；根据性能要求信息，分别确定与性能要求信息对应的目标编码器、目标注意力组件和目标解码器；获得包括目标编码器、目标注意力组件和目标解码器的目标语音合成模型；通过目标语音合成模型将接收的字符信息合成为语音信息。

[0006] 另一方面，本申请实施例提供了一种语音合成装置，该装置包括信息获得模块、确定模块、模型获得模块和语音合成模块。其中，信息获得模块用于获得针对语音合成模型的性能要求信息。确定模块用于根据性能要求信息，分别确定与性能要求信息对应的目标编码器、目标注意力组件和目标解码器。模型获得模块用于获得包括目标编码器、目标注意力组件和目标解码器的目标语音合成模型。语音合成模块用于通过目标语音合成模型将接收的字符信息合成为语音信息。

[0007] 另一方面，本申请实施例提供了一种服务器，包括：一个或多个处理器；存储器；一个或多个程序，其中所述一个或多个程序被存储在所述存储器中并被配置为由所述一个或多个处理器执行，所述一个或多个程序配置用于执行上述的方法。

[0008] 另一方面，本申请实施例提供了一种计算机可读存储介质，其上存储有程序代码，该程序代码可被处理器调用执行上述的方法。

[0009] 本申请提供的方案，根据针对语音合成模型的性能要求信息，分别确定与该性能要求信息对应的目标编码器、目标注意力组件和目标解码器，并获得包括该目标编码器、目标注意力组件和目标解码器的目标语音合成模型，通过该语音合成模型将接收的字符信息合称为语音信息。如此，可以根据不同的性能要求灵活地获取语音合成模型，从而基于该语音合成模型提供符合该性能要求的语音合成服务，改善用户体验。

[0010] 本申请的这些方面或其他方面在以下实施例的描述中会更加简明易懂。

附图说明

[0011] 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。

[0012] 图1示出了本申请实施例提供的一种语音合成模型的架构示意图。

[0013] 图2为一种适用于本申请实施例的应用环境示意图。

[0014] 图3示出了本申请实施例提供的一种语音合成方法的流程示意图。

[0015] 图4示出了图3所示步骤S330的子步骤示意图。

[0016] 图5示出了本申请实施例提供的另一种语音合成模型的架构示意图。

[0017] 图6示出了图3所示实施例中的语音合成方法的另一流程示意图。

[0018] 图7示出了图3所示实施例中的语音合成方法的一个应用场景示意图。

[0019] 图8示出了本申请实施例提供的第一损失函数的建立流程。

[0020] 图9示出了本申请实施例提供的语音合成方法的另一流程示意图。

[0021] 图10A示出了本申请实施例提供的一种残差网络结构的示意图。

[0022] 图10B示出了本申请实施例提供的一种第二编码层的结构示意图。

[0023] 图11示出了本申请实施例提供的一种目标编码器的处理流程图。

[0024] 图12示出了本申请实施例提供的一种目标注意力组件的处理流程图。

[0025] 图13示出了本申请实施例提供的一种目标注意力组件的结构示意图。

[0026] 图14示出了本申请实施例提供的一种目标解码器的处理流程图。

[0027] 图15A示出了本申请实施例提供的一种目标解码器的结构示意图。

[0028] 图15B示出了本申请实施例提供的一种GRU层的结构示意图。

[0029] 图16示出了本申请实施例提供的一种目标语音合成模型的结构示意图。

[0030] 图17示出了本申请实施例提供的一种语音合成装置的框图。

[0031] 图18示出了本申请实施例的用于执行根据本申请实施例的语音合成方法的服务器的框图。

[0032] 图19示出了本申请实施例的用于保存或者携带实现根据本申请实施例的语音合成方法的程序代码的存储单元。

具体实施方式

[0033] 为了使本技术领域的人员更好地理解本申请方案，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述。

[0034] 一些实施方式中，基于深度学习的端到端语音合成系统是一个独立而完整的模型，即，其结构是一个黑盒子，通常难以调整。在此情况下，基于深度学习的端到端语音合成系统难以适应不同的性能要求。

[0035] 发明人经过长期的研究，提出一种语音合成方法、装置及服务器，可以根据不同的性能要求获得对应的语音合成模型，从而基于该语音合成模型提供符合相应性能要求的语

音合成服务。

[0036] 请参照图1,图1为本申请实施例提供的一种语音合成模型100的架构示意图。语音合成模型100可以包括多个组件,该多个组件比如可以是编码器(Encoder)101、注意力(Attention)组件102、解码器(Decoder)103及声码器(Vocoder)104。

[0037] 其中,编码器101用于接收输入的字符信息901,并提取该字符信息901的字符特征向量。这里的字符信息901可以是文本信息或文本信息的注音字符,比如,中文文本信息的注音字符可以是拼音。文本是指书面语言的表现形式,是指具有特定含义的一个或多个字符,例如可以是具有特定含义的字、词、短语、句子、段落或篇章,文本信息则可以理解为包含文本的信息。

[0038] 上述的字符信息901可以包括多个字符,在语音合成过程中,需要将每个字符合成为语音。其中,在将字符信息901中的一个目标字符x_t合成语音时,除了需要使用目标字符x_t的字符特征向量,可能还需要使用字符信息901中的其他字符(例如,字符x_1,x_2,x_N)的字符特征向量。在此情况下,需要确定字符信息901中每个字符的字符特征向量与目标字符x_t的关联程度,该关联程度可以理解为目标编码器输出的每个字符的字符特征向量与目标字符x_t对应的注意力得分(score)。注意力组件102可以用于学习该注意力得分,并输出至解码器103。

[0039] 解码器103用于根据编码器101输出的字符特征向量及注意力组件102输出的注意力得分来预测字符信息901中的每个字符对应的声学特征向量,并输出至声码器104。声码器104用于根据声学特征向量合成与字符信息901对应的语音信息。

[0040] 本申请实施例中,编码器101、注意力组件102、解码器103和声码器1004是独立建立的组件,它们通过拼接形成所述语音合成模型100。

[0041] 请参照图2,图2是一种适用于本申请实施例的应用环境示意图。其中,服务器200通过网络300与终端设备400通信连接。终端设备400可以安装有客户端410,并可以通过客户端410登录至服务器200,以使用服务器200提供的服务,如,语音合成服务。

[0042] 其中,服务器200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,还可以是提供云计算、大数据和人工智能平台等基础云计算服务的云服务器。终端设备400可以是智能手机、平板电脑、笔记本电脑、个人计算机(Personal Computer,PC)、便携式穿戴设备等。客户端410可以是语音合成应用程序或其他任意需要使用语音合成服务的应用程序,还可以是供开发人员访问并配置服务器200的应用程序。本申请实施例对此没有限制。

[0043] 请参照图3,图3为本申请实施例提供的一种语音合成方法的流程示意图,该方法可以应用于图2所示的服务器200。下面对该方法包括的步骤进行描述。

[0044] S310,获得针对语音合成模型的性能要求信息。

[0045] 其中,性能要求信息可以包括至少一个性能指标信息,性能指标信息是表示语音合成模型的任意一个模型性能的信息,比如可以是表示语音合成模型的算力成本大小的信息、表示语音合成模型适合处理的字符信息类型的信息、表示语音合成模型对不同字符信息的表征能力的信息等。可以理解,上述的性能指标信息仅为举例说明,而非用于限制本申请。比如,性能指标信息还可以是表示语音合成模型的并行能力的信息以及表示语音合成模型合成的语音信息与字符信息的匹配程度的信息等。

[0046] 本实施例中，性能要求信息的确定方式可以有多种。一种实施方式中，可以由用户在终端设备400的客户端410上输入，并由客户端410发送至服务器200。这里的用户可以是需要使用语音合成服务的用户，也可以是可对服务器200进行配置的开发人员。

[0047] 一个例子中，需要使用语音合成服务的用户可以在客户端410输入自己对语音合成模型的性能要求，客户端410可以从用户输入的信息中识别出性能指标信息，识别出的各个性能指标信息即所述性能要求信息，并将所述性能要求信息发送给服务器200。另一个例子中，开发人员例如可以通过线下方式获得用户对语音合成模型的性能要求信息，并通过客户端410输入获得的性能要求信息，以使客户端410将性能要求信息发送给服务器200。当然，开发人员也可以直接在服务器200输入其获得的性能要求信息，本实施例对此没有限制。

[0048] 另一种实施方式中，服务器200可以根据终端设备400所处的应用场景来确定客户端410对语音合成模型的性能要求信息。终端设备400可以对所处应用场景进行识别，并将识别结果发送给服务器200，服务器200可以根据该识别结果确定性能要求信息。

[0049] 一个例子中，该应用场景可以是终端设备400当前所处的物理环境，对应地，终端设备400对应用环境的识别结果可以是通过传感器采集的环境信息，例如噪声信息。服务器200可以根据噪声信息确定终端设备400所处环境的噪声大小，当噪声大小达到阈值时，可以确定性能要求信息包括：语音合成模型需要对合成的语音信息进行信号增强。此时，服务器200在S320中可以选取带有信号增强结构的解码器作为目标解码器。

[0050] 另一个例子中，应用场景可以理解成终端设备400对语音合成模型的调用方式。详细地，终端设备400可以确定本设备调用语音合成模型时所使用的接口的类型，并发送给服务器200。可以理解，接口的类型在此可以视为对应用场景的识别结果。服务器200在确定接口的类型是本地调用接口时，可以确定语音合成模型是在本地使用的，即集成于终端设备400的。由于终端设备的算力通常是有限的，可以确定性能要求信息包括：结构简单。

[0051] 服务器200如果确定接口的类型是远程调用接口，则可以确定语音合成模型是部署在服务器上的，则可以进一步识别该远程调用接口是否包括云端API(Application Programming Interface，应用程序编程接口)，如果是，可以确定语音合成模型是部署在云服务器上的。在此情况下，也可以确定性能要求信息包括：结构简单。

[0052] 进一步地，服务器200还可以针对一些相似的性能要求信息设置不同级别，比如，对于结构简单这一性能要求信息，可以设置至少两个级别，如第一级别和第二级别。其中，满足第一级别性能要求的语音合成模型的结构比满足第二级别要求的语音合成模型的结构更简单。对应地，服务器200确定性能要求信息表示第一级别的性能要求时，可以在S320中分别确定结构简单的编码器、注意力组件和解码器作为目标编码器、目标注意力组件和目标解码器。如果确定性能要求信息标识第二级别的性能要求，则通过S320确定的目标编码器、目标注意力组件和目标解码器只需要其中一者或两者结构简单即可。

[0053] 再一个例子中，应用场景可以理解成语音合成模型需要处理的字符信息的类型。对应地，终端设备400可以对采用语音合成模型的应用程序(如，客户端410)所处理的每一字符信息的类型进行统计，统计结果可以是目标时间段(如，1天、1个周或一个月等)内处理的每个类型的字符信息的数量。所述统计结果可以视为终端设备400对应用场景的识别结果，并可以被终端设备400发送给服务器200。

[0054] 一种方式中，服务器200在根据统计结果确定特定类型的字符信息的数量达到第一数量(如，100-500)时，可以确定性能要求信息是：语音合成模型需要适用于该特定类型的字符信息的识别。比如，该特定类型可以是超长字符信息，即，字符数量达到第二数量(如，50、70或100等)的字符信息。对应地，服务器200可以确定性能要求信息包括：语音合成模型需要适用于超长字符信息的识别。

[0055] 另一种方式中，服务器200在根据统计结果确定任意两种类型的字符信息的数量之间的差值小于第三数量时，可以确定语音合成模型需要能够良好地区分各种字符信息，因此，可以确定性能要求信息包括：语音合成模型对字符信息的表征能力良好。

[0056] 值得说明的是，上述的确定性能要求信息的方式仅为举例，本申请实施例还可以通过其他方式来确定针对语音合成模型的性能要求信息。

[0057] S320，根据所述性能要求信息，分别确定与所述性能要求信息对应的目标编码器、目标注意力组件和目标解码器。

[0058] 本实施例中，语音合成模型按照功能的不同可以被划分为多个组件，例如，上述的语音合成模型100中的编码器101、注意力组件102、解码器103和声码器104。针对编码器、注意力组件和解码器这三个组件中的每一者，可以分别建立至少一种结构的该组件。比如，可以建立至少一种结构的编码器、至少一种结构的注意力组件和至少一种结构的解码器。每种结构的组件与不同的性能指标信息相对应。建立的各种结构的组件可以存储于服务器200中，也可以存储于可供服务器200访问的其他服务器中，本实施例对此没有限制。

[0059] 服务器200可以存储有不同结构的组件与不同的性能指标信息之间的对应关系，这里的对应关系可以理解为数据记录，该数据记录可以包括至少一种结构的组件的标识和至少一性能指标信息。值得说明的是，同一组件的结构不同，其标识也不同。比如，如果两个编码器具有不同的结构，这两个编码器的标识也将不同。

[0060] 实施过程中，针对获得的性能要求信息中的每个性能指标信息，服务器200可以查找包含该性能指标信息的对应关系作为目标对应关系，则可以确定目标对应关系中的标识所指示的组件具有该性能指标信息对应的结构，换句话说，，可以将目标对应关系中的标识所指示的组件确定为该性能指标信息对应的组件。可以理解，如果目标对应关系包括某一编码器的标识，则该编码器可以视为S320中的目标编码器。如果目标对应关系包括某一注意力组件的标识，则该注意力组件可以视为S320中的目标注意力组件。如果目标对应关系包括解码器的标识，则该解码器可以视为S320中的目标解码器。

[0061] 一些情况下，服务器200根据获得的性能要求信息中的性能指标信息只可以确定目标编码器、目标注意力组件和目标解码器中的一部分(即，一者或两者)。比如，服务器200获得的性能要求信息包括两个性能指标信息c1和c2，并根据c1确定了目标编码器e1，根据c2确定了目标注意力组件a1，而没有确定解码器，这表示用户对于解码器相关的性能没有要求。在此情况下，一种实施方式中，可以从存储的各种结构的解码器中随机确定一者作为目标解码器。另一种实施方式中，服务器200可以记录有每种结构的组件通过S320被确定的次数，即命中频次，从而可以从存储的各种结构的解码器中选择命中频次最多的一者作为目标解码器。

[0062] 又比如，服务器200获得的性能要求信息只包括一个性能指标信息c3，并根据c3确定了目标编码器e2和目标解码器d1，而没有确定解码器，则解码器也可以通过上述的实施

方式确定。可以理解，在其他示例中，当无法确定注意力组件或编码器时，也可以通过上述的实施方式来确定目标注意力组件或目标编码器。

[0063] S330，获得包括所述目标编码器、所述目标注意力组件和所述目标解码器的目标语音合成模型。

[0064] 在确定目标编码器、目标注意力组件和目标解码器之后，可以通过图4所示的流程实现S330。

[0065] S331，确定语音合成模型的目标框架。

[0066] S332，按照目标框架对目标编码器、目标注意力组件和目标解码器进行组合，从而得到目标语音合成模型。

[0067] 本实施例中，按照目标框架组合而成的目标语音合成模型，可以按照目标框架对应的处理流程，将输入的字符信息合成为语音信息。

[0068] 一种实施方式中，目标框架可以是固定的框架，例如可以是图1所示的端到端语音合成模型的通用框架，在该通用框架中，目标编码器、目标注意力组件、目标解码器及预先配置的声码器可以依次连接，从而得到拼接的目标语音合成模型。

[0069] 另一种实施方式中，目标框架可以是动态框架，具体可以根据目标编码器、目标注意力组件和目标解码器的结构而定。例如，一些情况下，目标框架可以是图1所示的通用框架。另一些情况下，目标框架则可以是图5所示的框架。在图5所示的框架中，编码器501和注意力组件502可以是并行的两个组件，两者并行地连接至解码器503，解码器503再进一步与声码器504连接。

[0070] 一种实施方式中，编码器501和注意力组件502可以是完全并行，例如，编码器501用于将输入的字符信息编码为字符特征向量，而注意力组件502也可以基于字符信息本身来确定字符信息中每个字符对应的声学特征信息与字符信息中各个字符的关联程度，即注意力得分。

[0071] 另一种实施方式中，编码器501和注意力组件502可以是部分并行，例如，在在编码器501包括多个处理层的情况下，注意力组件502可以基于其中一些处理层的输出来确定上述的注意力得分。

[0072] 可以理解，上述的目标框架仅为示例，本实施例中，随着语音合成模型的各个组件的结构的不同，还可以采用其他框架来拼接目标编码器、目标注意力组件、目标解码器，在此情况下，拼接而成的目标语音合成模型可以按照该框架对应的处理流程，将字符信息合成为语音信息。

[0073] S340，通过所述目标语音合成模型将接收的字符信息合成语音信息。

[0074] 本实施例中，服务器200还可以获得针对语音合成模型的部署位置信息，从而可以在获得目标语音合成模型之后，将目标语音合成模型部署于所述部署位置信息指示的设备上。其中，用于部署目标语音合成模型的设备可以是服务器，也可以是终端设备，本实施例对此没有限制。

[0075] 当目标语音合成模型被部署于相应设备之后，可以接收字符信息并将该字符信息作为目标语音合成模型的输入信息，从而获得目标语音合成模型输出的语音信息，输出的语音信息即为所述字符信息对应的语音信息。值得说明的是，这里的字符信息可以是文本信息或文本信息对应的注音字符，本实施例对此没有限制。

[0076] 通过图3所示流程，可以根据不同的性能要求灵活地进行组件拼接，以形成符合该性能要求的目标语音合成模型，从而基于目标语音合成模型提供符合该性能要求的语音合成服务，换言之，可以针对不同场景定制化设计相应的语音合成模型，从而应对不同的语音合成服务需求。

[0077] 请再次参阅图3，下面将对图3所示的流程做进一步的详细阐述。

[0078] 本实施例中，语音合成系统的同一组件，结构不同时，其所需的输入和得到的输出可能存在差异。比如，不同结构的编码器需要的输入和得到的输出可能是不同的，不同结构的注意力组件需要的输入和得到的输出可能是不同的，不同结构的解码器需要的输入和得到的输出可能是不同的。而通过S320确定的目标编码器、目标注意力组件和目标解码器是基于获得的性能要求信息确定的，性能要求信息又是随机的，因此，用于拼接形成目标语音合成系统的各个组件也存在随机性，即难以预先确定。在此情况下，目标编码器、目标注意力组件、目标解码器、声码器中，任意两个相邻组件的输出和输入可能不匹配，比如目标编码器的输出和目标注意力组件的输入可能不匹配，目标注意力组件的输出和目标解码器的输入可能不匹配，目标解码器的输出和声码器的输入可能不匹配。

[0079] 针对上述问题，可以对建立的各种结构的组件，可以对它们的输入和输出进行归一化处理，以使同一组件在采用不同结构的情况下，具有相同的维度输入和相同维度的输出，且任意两个相邻组件中前一组件的输出与后一组件的输入具有相同的维度。换言之，各种结构的组件具有统一适配的接口，如此，通过S320确定的目标编码器、目标注意力组件和目标解码器可以与预先配置的声码器拼接成所述目标语音合成模型。

[0080] 本实施例中，在将目标语音合成模型部署于相应设备之前，可以对目标语音合成模型进行训练。基于上述对语音合成模型100的描述可知，目标语音合成模型执行的主要处理是：基于输入的字符信息预测语音信息的声学特征。因此，例如可以通过下文描述的训练过程实现对目标语音合成模型的训练：

[0081] 首先是样本采集阶段，可以采集字符信息和字符信息对应的真实语音信息，并从真实语音信息提取声学特征，提取的声学特征为真实声学特征。将字符信息和提取的真实声学特征作为一个样本，并将样本添加到训练数据集中。重复前述过程可以建立包括多个样本的训练数据集。

[0082] 然后是训练阶段，将每个样本中的字符信息输入目标语音合成模型，则目标语音合成模型可以输出预测声学特征，并可以通过一损失函数对预测声学特征和该样本中的真实声学特征进行计算，以得到所述损失函数的函数值，即预测声学特征与该样本中的真实声学特征之间的损失。基于所述损失对目标语音合成模型中的模型参数进行调整，以使所述损失函数满足优化条件。这里的优化条件例如可以是损失函数的函数值收敛，或是迭代次数达到设定的次数。

[0083] 为了便于与后文描述的其他损失函数区分开，本实施例将目标语音合成的损失函数描述为第一损失函数。本实施例中，第一损失函数可以灵活设置，例如可以是负数对数似然函数、均方误差(Mean Square Error，MSE)损失函数、交叉熵损失函数、连接时间分类(Connectionist Temporal Classification，CTC)损失函数等。本实施例对此没有限制。

[0084] 一些实施方式中，训练过程还可以包括测试阶段。其中，测试数据集的建立方式与上述的训练数据集的建立方式类似，在此不再赘述。实施时，可以按照训练阶段的处理方

式，将测试数据集内的每个样本中的字符信息输入经过训练的目标语音合成模型，从而得到输出的预测声学特征，并可以通过所述损失函数计算预测声学特征与该样本中的真实声学特征之间的损失是否符合目标条件(例如小于目标阈值等)，如果符合目标条件，则可以确定本次预测结果是准确的。在测试过程中，统计预测结果准确的次数，并根据该次数和测试次数计算目标语音合成模型的预测准确率，当预测准确率符合要求时，可以确定训练过程结束，从而可以将当前的目标语音合成模型部署于相应的设备，以用于提供语音合成服务。

[0085] 一些场景中，用户对于合成的语音的音色有要求。在此情况下，上述的训练过程可以基于符合用户的音色要求的声音数据实现。换句话说，在执行S340之前，本实施例提供的语音合成方法还可以包括图6所示的流程。

[0086] S610，获得针对语音合成模型的音色要求信息。

[0087] 其中，音色要求信息的输入方式与性能要求信息的类似，可以由用户在客户端410输入并被客户端410发送给服务器200，或是直接通过服务器200输入。音色要求信息可以是描述音色类型的信息，这里的音色类型可以包括情感类型、发音人性别、发音人年龄等，本实施例对此没有限制。

[0088] S620，根据所述音色要求信息获取声音数据。

[0089] 实施过程中，在获得音色要求信息之后，在训练过程的样本采集阶段，针对每一字符信息可以采集符合该音色要求信息的真实语音信息，再基于该真实语音信息提取真实声学特征，所述真实声学特征可以理解为S620中的声音数据。

[0090] S630，基于所述声音数据对所述目标语音合成模型进行模型训练，使所述目标语音合成模型的第一损失函数达到优化条件。

[0091] 实施过程中，在基于音色要求信息获得声音数据后，可以基于所述声音数据执行上文描述的训练阶段对应的流程，以使第一损失函数满足优化条件。如此，经过训练的目标语音合成模型可以更加适用于符合所述音色要求信息的语音信息的合成，从而可以进一步改善用户体验。

[0092] 请参照图7，在一个具体的应用场景中，基于本申请实施例提供的语音合成方法，服务器200可以获得需要提供语音合成服务的用户U1对语音合成模型的性能要求信息和音色要求信息，并按照图3所示的S310-S330，定制出与性能要求信息对应的目标语音合成模型，在经过户U1确认的情况下，目标语音合成模型可以上线，也即可以被部署于相应的设备，从而对外(即，对需要使用语音合成服务的用户U2)提供语音合成服务，也即，接收字符信息，并将接收的字符信息合成语音。

[0093] 可选地，本实施例中，第一损失函数除了可以直接设置为上文提及的损失函数，还可以通过图8所示的流程建立。详细描述如下。

[0094] S810，获取目标编码器在各编码次序输出的目标字符向量各自在当前解码次序的注意力得分，得到一注意力得分序列。

[0095] 请再次参照图1，其中示出了可用于输入目标语音合成模型的字符信息901的内容结构示意图。字符信息901包括N个字符，其中，第t个字符表示为x_t，t为范围[1，N]内的整数，N为大于1的正整数。例如，第1个字符表示为x_1，第2个字符表示为x_2，第N个字符表示为x_N。

[0096] 如果将字符信息901作为目标语音合成模型的输入，则字符信息901中的各个字符将会按照排列顺序依次被输入目标编码器，对应地，目标编码器可以按照各个字符的接收顺序依次将每个字符编码为字符特征向量并将该字符特征向量输出。如此，目标编码器基于字符信息可以输出N个字符特征向量。其中，字符信息901中各个字符的排列顺序就是目标编码器接收所述各个字符的顺序，也是目标编码器针对所述各个字符的编码次序。目标编码器基于每个字符编码并输出的字符特征向量即为S810中描述的目标字符特征向量。

[0097] 对应地，目标解码器需要按照字符信息901中各个字符的排列顺序依次预测每个字符的声学特征向量，预测声学特征向量的过程即为解码过程。

[0098] 参照上文关于图1所示注意力组件102的描述，目标解码器在对每个字符进行解码时，以当前被解码字符是x_t为例，需要按照字符信息901中各个字符(包括字符x_t本身)的字符特征向量与字符x_t的关联程度来基于所述各个字符的字符特征向量预测字符x_t的声学特征向量。换句话说，针对每个解码次序，目标注意力组件会计算字符信息901中每个字符的字符特征向量与当前解码次序的字符(即，当前需要被解码的字符，如x_t)之间的关联程度，该关联程度即为注意力得分。该关联程度也可以理解为在每个解码次序需要关注字符信息901中的哪些字符。

[0099] 实施过程中，针对每个解码次序可以求解出N个注意力得分，N个注意力得分与目标编码器输出的N个字符特征向量依次对应。这N个注意力得分可以形成一个序列，即所述注意力得分序列。

[0100] S820，确定所述注意力得分序列的熵。

[0101] S830，将所述熵叠加至第二损失函数，得到第一损失函数。

[0102] 其中，可以通过序列的信息熵的计算式来计算每个注意力得分序列的熵。该计算式例如可以是：$H(x)=-\sum p(x_i)\log(p(x_i))$，其中，$i=1,2,\cdots,N$。第二损失函数可以是预先设置的损失函数，比如可以是上文提及的负数对数似然函数、MSE损失函数、交叉熵损失函数、CTC损失函数等中的任意一者，或者也可以是其他的损失函数。

[0103] 本实施例中，在每个解码次序计算出一个注意力得分序列之后，即可通过S820计算该注意力得分序列的熵，并将该熵叠加至第二损失函数上。如此，在将各个解码次序求得的注意力得分序列的熵都叠加至第二损失函数后，即可得到第一损失函数。换言之，第一损失函数是第二损失函数与N个注意力得分序列的熵之和。

[0104] 将基于图8所示流程确定的第一损失函数作为目标语音合成模型的损失函数，在训练阶段，除了需要最小化预测声学特征与真实声学特征之间的损失，还需要最小化注意力得分序列的熵，从而可以降低每个解码次序求得的注意力得分序列的不确定性，也即降低每个解码次序需要关注的字符的不确定性，从而可以增强模型训练的准确性，提高目标语音合成模型的鲁棒性。

[0105] 进一步地，上述的N个注意力得分序列的熵可以按照目标权重叠加至第二损失函数，换言之，可以将每个注意力得分序列的熵与目标权重的乘积叠加至第二损失函数。其中，目标权重可以具有预设的初始值，在训练过程中，目标权重可以随着迭代次数的增加而逐步增大。如此，可以避免一开始对注意力熵的限制过大而导致目标语音合成模型无法正常地对齐。

[0106] 在图3所示的S320中，随着性能要求信息的不同，确定的目标编码器、目标注意力

组件和目标解码器也不尽相同。鉴于这种灵活确定所需组件以组合成目标语音合成模型的处理流程，为了适应更多的需求，可以进一步基于不同的性能指标，对语音合成模型的不同组件的结构进行精细化的设计和改进，并对结构改进后的组件及该组件与性能指标信息的对应关系进行存储，以供具有该性能要求的用户选用。

[0107] 一个例子中，服务器200可以存储第一性能指标信息和具有残差网络结构(ResNet)的编码器之间的对应关系。在此情况下，如图9所示，S320可以包括步骤S321。

[0108] S321，若性能要求信息包括第一性能指标信息，将具有残差网络结构(ResNet)的编码器确定为目标编码器。

[0109] 其中，第一性能指标信息可以是表示对不同字符信息的表征能力良好的信息，也可以是表示对不同字符信息的区分能力好的信息。

[0110] 实施过程中，服务器200可以通过语义识别模型来识别性能要求信息所属的语义类型。详细地，将性能要求信息输入语义识别模型之后，语义识别模型可以输出该性能要求信息属于各语义类型的概率。所述各语义类型与服务器200存储的对应关系中出现的性能指标信息一一对应。例如，服务器200存储了K个性能指标信息与不同结构的组件之间的对应关系，则语义识别模型针对输入信息可以输出该输入信息分别属于K个语义类型的概率，K个语义类型与K个性能指标信息一一对应。

[0111] 如此，对于输入的性能要求信息，当语义识别模型输出的最大概率对应的语义类型是第一性能指标信息对应的语义类型时，可以将第一性能指标信息对应的标识所指示的组件，也就是具有残差网络结构的编码器，确定为目标编码器。

[0112] 请参照图10A，其中示例性地示出了残差网络结构1000的结构示意图。残差网络结构包括两个编码层，分别为第一编码层L1和第二编码层L2。第一编码层L1的输出信息被叠加至第二编码层L2的输出信息。换言之，目标编码器的输出为第一编码层L1的输出信息和第二编码层L2的输出信息叠加而成的信息。

[0113] 其中，第一编码层L1用于将字符信息转换为向量，例如可以通过词嵌入(Word embedding)算法或词向量(Word2vec)算法实现。第二编码层L2可以为用于处理序列信息的网络单元，例如可以是循环神经网络(Recurrent Neural Network，RNN)单元，比如长短时记忆网络(Long Short-Term Memory，LSTM)、循环门单元(Gate Recurrent Unit，GRU)等。

[0114] 对应地，在目标编码器是具有残差网络结构的编码器时，目标编码器可以通过图11所示的步骤对字符信息进行处理，以将字符信息编码为字符特征向量。

[0115] S1101，通过第一编码层，按照接收次序对所述字符信息中的每个字符进行编码，得到第一字符特征向量。

[0116] S1102，通过第二编码层处理所述第一字符特征向量，得到第二字符特征向量。

[0117] 实施过程中，字符信息中的各个字符按照排列次序依次被输入第一编码层L1，第一编码层L1可以按照接收次序依次将每个字符转换为向量，该向量可以为第一字符特征向量。第一字符特征向量将继续被输入至第二编码层L2，第二编码层L2可以对第一字符特征向量进行编码，从而输出第二字符特征向量。其中，第一字符特征向量可以理解为字符的浅层特征向量，第二字符特征向量可以理解为字符的高层特征向量。

[0118] S1103，拼接(concat)所述第一字符特征向量和所述第二字符特征向量，得到当前编码次序的字符对应的目标字符特征向量。

[0119] 第一编码层L1输出的第一字符特征向量将被添加至第二编码层L2的输出端，从而与第二编码层L2输出的第二字符特征向量拼接在一起。第一字符特征向量与第二字符特征向量的拼接结果即为所述目标字符特征向量，也即目标编码器的输出。

[0120] 以当前编码次序得到的第一字符特征向量是x为例，第二编码层L2输出的第二字符特征向量可以表示为H(x)，而目标编码器输出的目标字符特征向量可以表示为F(x)，前述三者的关系为F(x)＝H(x)+x。在针对目标语音合成模型的训练阶段，可以调整模型参数，以使F(x)尽可能地接近x，也即H(x)＝F(x)-x尽可能地接近0。如此，在经过训练的目标语音合成模型中，第二编码层L2输出的第二字符特征向量实际表征的是目标编码器的输出F(x)与第一编码层L1的输出x之间的残差信息，残差信息属于细化的编码特征，而第一字符特征向量x作为浅层特征向量是比较粗略的编码特征。换句话说，目标编码器的输出F(x)包括残差信息和字符的浅层特征向量，通过粗略编码和细化编码相结合的方式，使得从字符信息提取的字符特征向量具有更强的表征能力，有助于后续目标注意力组件的处理。

[0121] 请参照图10B，其中以LSTM单元为例示出了第二编码层L2的结构示意图。其中，δ表示表示Sigmoid激活函数，tanh表示tanhh激活函数，⊗表示按元素相乘，⊕ 表示相加。第二编码层L2包括多个门控单元，分别为输入门、遗忘门和输出门。此外，还具有细胞状态参数(Cell State)，用于记录第二编码层L2在各个编码次序的状态信息。

[0122] 其中，y_t表示编码次序为t的字符经第一编码层L1处理得到的第一字符特征向量。h_{t-1}表示第二编码层L2在编码次序t-1的隐藏状态，h_t表示第二编码层L2在编码次序t的隐藏状态，C_{t-1}表示第二编码层L2在编码次序t-1的细胞状态，C_t表示第二编码层L2在编码次序t的细胞状态。在第二编码层L2中，遗忘门在编码次序t的输出f_t可以通过如下计算式计算：

[0123] $f_t=\delta(W_{yf}y_t+W_{hf}h_{t-1}+W_{cf}C_{t-1}+b_f)$；

[0124] 其中，遗忘门的输出f_t通常属于[0,1]区间，用于表示对第二编码层L2在当前编码次序对上一编码次序的细胞状态参数的遗忘比例。W_{yf}、W_{hf}、W_{cf}为权重矩阵，b_f为偏置矩阵，可以通过训练得到。

[0125] 第二编码单元L2的输入门可以根据输入的y_t和h_{t-1}对C_{t-1}进行处理，以得到新的细胞状态参数C_t，也即第二编码层L2在当前编码次序t的细胞状态参数。详细地，输入门在编码次序t的输出值i_t可以通过如下计算式计算：

[0126] $i_t=\delta(W_{yi}y_t+W_{hi}h_{t-1}+W_{ci}C_{t-1}+b_i)$；

[0127] 对应地，第二编码层L2在编码次序t的细胞状态参数C_t可以通过以下计算式计算：

[0128] $C_t=f_tC_{t-1}+i_t\tanh(W_{yt}y_t+W_{hc}h_{t-1}+b_c)$；

[0129] 其中，W_{yi}、W_{hi}、W_{ci}、W_{yt}、W_{hc}为权重矩阵，b_i、b_c为偏置向量，可以通过训练确定。

[0130] 第二编码单元L2的输出门在编码次序t的输出O_t可以通过以下计算式得到：

[0131] $O_t=\delta(W_{yo}y_t+W_{ho}h_t+W_{co}C_t+b_o)$；

[0132] 其中，W_{yo}、W_{ho}、W_{co}为权重矩阵，b_o为偏置向量，均可以通过模型训练确定。第二编码层L2在编码次序t的隐藏状态，则可以通过如下计算式确定：

[0133] $h_t=O_t\tanh(C_t)$。

[0134] 第二编码层L2在编码次序t输出的第二字符特征向量则可以为$h_tW_{pre}+b_{pre}$，其中，W_{pre}为权重矩阵，b_{pre}为偏置向量，均可以通过模型训练确定。

[0135] 可以理解,在目标编码器具有残差网络结构的情况下,在S340的执行过程中,对于字符信息的编码过程可以和图11所示流程类似。当然,上述的具有残差网络结构的编码器仅为举例,当性能要求信息发生变化时,目标编码器还可以是其他结构的编码器。

[0136] 另一个例子中,服务器200可以存储第二性能指标信息和包括参数拟合层的注意力组件之间的对应关系。在此情况下,请再次参照图9,S320可以包括步骤S322。

[0137] S322,若性能要求信息包括第二性能指标信息,将包括参数拟合层的注意力组件确定为目标注意力组件。

[0138] 其中,第二性能指标信息可以是表示能够稳定识别超长字符信息的信息。实施过程中,可以利用上述的语义识别模型对性能要求信息的语义类型进行识别,如果语义识别模型输出的最大概率对应的语义类型是与第二性能指标信息对应的语义类型时,可以将第二性能指标信息对应的标识所指示的组件(即,包括参数拟合层的注意力组件)确定为目标注意力组件。

[0139] 其中,参数拟合层用于拟合高斯混合模型的参数特征信息,这里的高斯混合模型是目标编码器在每个编码次序编码的字符与目标解码器在当前解码次序的字符的相关程度所服从的概率分布模型。换句话说,该高斯混合模型可以理解为:目标编码器在每个编码次序输出的字符特征向量在当前解码次序的注意力得分所服从的概率分布模型。在此情况下,包括所述参数拟合层的注意力组件(即,目标注意力组件)可以称为基于高斯混合模型的注意力(Gaussian Mixed Model-Attention,GMM-Attention)组件。其中,所述参数拟合层可以为神经网络,例如可以为深度神经网络(Deep Neural Network,DNN)。

[0140] 在目标注意力组件是上述的GMM-Attention组件的情况下,目标注意力组件可以通过图12所示的步骤S1201至S1202,确定高斯混合模型的模型参数。这里的模型参数包括高斯混合模型中每个高斯分布的权重、方差和均值。

[0141] S1201,通过参数拟合层处理目标编码器在各编码次序的隐藏状态以及目标解码器在当前解码次序的隐藏状态,得到高斯混合模型的第一参数特征信息、第二参数特征信息和第三参数特征信息。

[0142] S1202,通过软最大化(Softmax)函数处理所述第一参数特征信息,得到高斯混合模型中每个高斯分布的权重;根据所述第二参数特征信息得到所述高斯混合模型中每个高斯分布的方差;通过软加(Softplus)函数处理所述第三参数特征信息,得到所述高斯混合模型中每个高斯分布的均值。

[0143] 请参照图13,其中以参数拟合层是DNN为例,示例性地示出了GMM-Attention组件1100的结构示意图。下面将结合图13所示结构对图12所示流程进行介绍。

[0144] 基于上文对图10B所示的第二编码层L2的介绍可知,第二编码层L2在各个编码次序都具有对应的隐藏状态,比如,第二编码层在编码次序t的隐藏状态为h_t。对应地,目标解码器在各个解码次序也具有对应的隐藏状态。

[0145] 实施过程中,目标解码器在预测字符信息中每个字符的声学特征向量,即进行每个解码次序的解码处理时,可以通过目标注意力组件(如,GMM-Attention组件1300)首先获取目标编码器在各个编码次序的隐藏状态以及目标解码器在当前解码次序的隐藏状态,并将获取的各隐藏状态输入DNN,DNN可以输出三个序列m1、m2和m3。

[0146] 本实施例中,GMM-Attention组件1300可以通过高斯混合模型的概率分布函数来

计算目标编码器在每个编码次序输出的目标字符特征向量在当前解码次序的注意力得分。可以理解，高斯混合模型是多个高斯分布加权求和的结果。本实施例中，高斯混合模型包括的高斯分布的数量可以和字符信息包括的字符数量相同。对应地，该高斯混合模型的概率分布函数可以通过以下计算式表示：

[0147] $$P(y|\theta)=\sum_{t=1}^{N}\alpha_t\phi(y|\theta_t),\ t\epsilon[1,N]$$

[0148] 其中，y表示高斯混合模型的变量，可以理解为目标解码器在当前解码次序的声学特征向量。$\phi(y|\theta_t)$可以表示第t个高斯分布的高斯分布密度函数，其中，$\theta_t=(\mu_t,\sigma_t^2)$，$\mu_t$为第t个高斯分布的均值，$\sigma_t^2$为第t个高斯分布的方差，$\alpha_t$可以表示第t个高斯分布的权重。

[0149] 本实施例中，上述的m1可以视为S330-4中的第一参数特征信息，可以用于求解上述概率分布函数中各高斯分布的权重。详细地，可以通过软最大化函数对第一参数特征信息(即序列m1)进行处理，所述软最大化函数的输出即为高斯混合模型中的各高斯分布的权重α_t组成的权重序列。

[0150] 上述的m2可以视为S330-4中的第二参数特征信息，可以用于求解高斯混合模型中每个高斯分布的方差。示例性地，可以通过指数函数对第二参数特征信息(如，序列m2)进行处理，所述指数函数的输出即为高斯混合模型中的各高斯分布的方差σ_t^2组成的方差序列。

[0151] m3可以视为S330-4中的第三参数特征信息，可以用于求解高斯混合模型中每个高斯分布的均值。详细地，可以通过软加函数对第三参数特征信息(如，序列m3)进行处理，所述软加函数的输出即为高斯混合模型中各高斯分布的均值μ_t组成的均值序列。

[0152] 上述流程中，通过软最大化函数来求解高斯分布的权重、通过软加函数来求解高斯分布的均值，由于软最大化函数和软加函数上升趋势相对比较平缓，降低了模型训练过程中出现梯度爆炸的概率，从而使得目标语音合成模型的训练过程可以更为稳定。

[0153] 进一步地，目标注意力组件可以通过图12所示的步骤S1203至S1205处理目标编码器输出的目标字符特征向量。详细描述如下。

[0154] S1203，根据所述高斯混合模型中每个高斯分布的权重、方差和均值，得到所述高斯混合模型在当前解码次序的概率分布函数。

[0155] S1204，根据所述概率分布函数和所述目标编码器在每个编码次序输出的目标字符特征向量，得到该目标字符特征向量在当前解码次序的注意力得分。

[0156] 在确定高斯混合模型中每个高斯分布的权重、方差和均值之后，即可确定上述的概率分布函数，并将目标编码器分别在每个编码次序输出的目标字符特征向量作为概率分布函数的输入，从而可以得到该目标字符特征向量在当前解码次序的注意力得分。通过所确定的概率分布函数对目标编码器在各个编码次序输出的目标字符特征向量分别进行处理，可以得到当前解码次序的注意力得分序列，注意力得分序列中的注意力得分与目标编码器在各编码次序输出的目标字符特征向量依次对应。

[0157] S1205，根据所述目标编码器在各编码次序输出的目标字符向量及每个目标字符向量在当前解码次序的注意力得分，得到所述目标解码器在当前解码次序的注意力向量。

[0158] 实施过程中，可以通过归一化处理将得到的注意力得分映射为[0,1]区间内的值，即注意力权重。如此，可以将当前解码次序的注意力得分序列转换为注意力权重序列，并可以将当前解码次序的注意权重序列与目标编码器输出的各目标字符向量按次序对应相乘，再将得到的N个乘积相加，得到的结果即为所述目标解码器在当前解码次序的注意力向量。

一些场景中,注意力向量也可以称为上下文向量(Context vector)。该上下文向量可以作为目标解码器的输入。

[0159] 另一个例子中,服务器200可以存储有第三性能指标信息和包括依次连接的循环门单元(Gate Recurrent Unit,GRU)和LSTM层的解码器之间的对应关系。在此情况下,请再次参照图9,S320可以包括步骤S323。

[0160] S323,若性能要求信息包括第三性能指标信息,将包括依次连接的GRU层和LSTM层的解码器确定为目标解码器。

[0161] 其中,第三性能指标信息可以是表示算力成本低的信息。实施过程中,可以利用上文描述的语义识别模型对性能要求信息进行语义识别,如果语义识别模型确定性能要求信息的最大概率对应的语义类型是与第三性能指标信息对应的语义类型时,可以将第三性能指标信息对应的标识所指示的组件(即,包括依次连接的GRU层和LSTM层的解码器)确定为目标解码器。

[0162] 在此情况下,目标解码器可以通过图14所示的步骤获得字符信息对应的预测声学特征信息。详细描述如下。

[0163] S1401,获取当前解码次序的注意力向量及所述当前解码次序的前一解码次序的注意力向量。

[0164] 其中,当前解码次序的注意力向量和前一解码次序的注意力向量均可以通过S1205处理得到。

[0165] S1402,通过GRU层,对所述前一解码次序的注意力向量、所述当前解码次序的目标字符特征向量及所述GRU层在前一解码次序的解码信息进行处理,得到第一声学特征向量。

[0166] S1403,通过LSTM层处理所述第一声学特征及所述当前解码次序的注意力向量,得到第二声学特征向量。

[0167] 请参照图15A,其中示例性地示出了包括GRU层和LSTM层的目标解码器1300的结构示意图。下面结合图15A对图14所示处理流程进行阐述。

[0168] 本实施例中,目标解码器1000可以包括依次连接的预处理网络(PreNet)层L3、GRU层L4、LSTM层L5、全连接(Fully Connected,FC)层L6以及后处理网络(PostNet)层L7。

[0169] 示例性地,GRU层L4的结构可以如图15B所示。其中,δ表示表示Sigmoid激活函数,tanh表示tanhh激活函数,⊗表示按元素相乘,⊕ 表示相加。GRU层L4包括两个门控单元,分别为更新门和重置门。y_t表示GRU层L4在第t个解码次序的输入向量,本实施例中,y_t可以根据第t-1个解码次序的注意力向量、目标编码器在第t个编码次序输出的目标字符向量及目标解码器1000在第t-1个解码次序的解码信息得到。

[0170] 上述的解码信息例如可以是目标解码器在第t-1个解码次序输出的预测声学特征信息(如,预测频谱信息)。示例性地,PreNet层L3可以用于将该预测声学特征信息转换为预测声学特征向量。如此,第t-1个解码次序的注意力向量、目标编码器在第t个编码次序输出的目标字符向量及目标解码器1000在第t-1个解码次序的预测声学特征向量可以拼接成y_t。

[0171] GRU层L4中,h_{t-1}表示GRU层L4在解码次序t-1的隐藏状态,h_t表示GRU层L4在解码次序t-1的隐藏状态,$\tilde{h}_t$表示当前解码次序的候选隐藏状态。

[0172] 在GRU层L4中,重置门在解码次序t的输出r_t可以通过以下计算式计算:

[0173] $r_t=\delta(W_\delta[h_t,y_t])$，

[0174] 本实施例中，[ ]表示两个向量相连接。W_δ为权重矩阵，可以通过模型训练过程确定。r_t用于表示需要忽略的前一解码次序的状态信息的多少，r_t越大，表示需要忽略的前一解码次序的状态信息越多。

[0175] 更新门的输出z_t可以通过以下计算式计算：

[0176] $z_t=\delta(W_z\cdot[h_t,y_t])$，

[0177] 本实施例中，W_z为权重矩阵，可通过模型训练过程确定。z_t表示需要使用的前一解码次序的状态信息的多少，z_t越大，表示需要使用的前一解码次序的状态信息越多。

[0178] 候选隐藏状态$\tilde{h}_t$可以通过以下计算式计算：

[0179] $$\tilde{h}_t=\tanh(W_{\tilde{h}}\cdot[r_t*h_t,y_t]),$$

[0180] 本实施例中，$W_{\tilde{h}}$为可通过模型训练确定的权重矩阵，*表示矩阵元素相乘。

[0181] 本实施例中，目标解码器在当前解码次序的隐藏状态可以是：GRU层L4在当前解码次序的隐藏状态h_t，其中，GRU层L4在当前解码次序的隐藏状态h_t可以通过以下计算式计算：

[0182] $$h_t=(1-z_t)*h_{t-1}+z_t*\tilde{h}_t。$$

[0183] GRU层L4基于第t个解码次序的输入向量y_t和GRU层L4在第t-1个解码次序的隐藏状态h_{t-1}输出的第一声学特征向量可以通过这一计算式得到：$\delta(W_o\cdot h_t)$。其中，W_o是可通过模型训练确定的权重矩阵。

[0184] 本实施例中，LSTM层L5的结构与图10B所示的第二编码层类似。不同之处在于，LSTM层L5的输入向量为当前解码次序(如，第t个解码次序)的注意力向量以及GRU层L4输出的第一声学特征向量拼接而成的向量。LSTM层L5的输出即为第二声学特征向量。

[0185] FC层L6用于对提取的第二声学特征向量进行整合，其输入可以是当前解码次序(如，第t个解码次序)的注意力向量和第二声学特征向量拼接而成的向量。

[0186] 通过图15A所示的目标解码器1300相较于采用多个LSTM层，减少了参数数量，且通过在每一层引入注意力向量(即，上下文向量)弥补因参数减少带来的信息损失，如此可以实现适用于低算力成本的场景的语音合成模型。

[0187] 进一步地，在得到整合后的第二声学特征向量之后，S330还可以包括步骤S330-12。

[0188] S1404，根据第二声学特征向量得到并输出目标解码器在所述当前解码次序的预测声学特征信息。

[0189] 这里的第二声学特征向量可以是通过FC层L6整合后的第二声学特征向量。输出的各个预测声学特征信息可以通过PostNet层L7进行串联，如此，可以提高后续的语音信息的生成质量。

[0190] 在得到预测声学特征信息后，可以进一步通过声码器，将目标解码器在各解码次序输出的预测声学特征信息合成为语音信息。

[0191] 一个例子中，声码器处理的可以是经过PostNet层L7的后处理的预测声学特征信息。详细地，声码器可以是Griffin-Lim声码器或者Wavenet声码器，本实施例对此没有限制。

[0192] 在其他的一些例子中，服务器200还可以存储其他性能指标信息与不同结构的组件之间的对应关系。比如，第四性能指标信息与变形器(Transformer)的编码器结构和解码器结构之间的对应关系，其中第四性能指标信息可以是表示并行能力良好的信息。又比如，第五性能指标信息与时长预测器之间的对应关系，采用时长预测器作注意力组件，可以预测每个字符的发音时长，从而可以对编码器输出的每个字符的目标字符特征向量进行复制扩张，以使复制扩张后的目标字符向量与该字符的发音时长匹配，也即，字符特征和声学特征具有较高的匹配程度。基于此，第五性能指标信息可以是表示字符特征与声学特征的匹配度高的信息。

[0193] 可以理解，实际应用中，通过S310获得的性能要求信息可以包括多个性能指标信息。比如，语义识别模型针对输入的性能要求信息输出的概率中，存在多个概率值比较接近，如差值小于0.01，在此情况下，可以确定所述多个概率值分别对应的多个语义类型，进而确定所述多个语义类型分别对应的多个性能指标信息，所确定的多个性能指标信息就是性能要求信息包括的性能指标信息。

[0194] 示例性地，如果确定性能要求信息同时包括上述的第一性能指标信息、第二性能指标信息、第三性能指标信息，则可以将图10A所示的残差网络结构1000确定为目标编码器，将图13所示的GMM-Attention组件1300确定为目标注意力组件，目标解码器则为图15A所示组件，将所确定的目标编码器、目标注意力组件和目标解码器依次拼接，可以得到图16所示的结构，该结构可以和声码器连接以形成目标语音合成模型。包括该结构的目标语音合成模型可以较好地区分不同字符信息，且可以适用于超长字符信息的语音合成，以及适用于低算力成本的应用场景，如需要将目标语音合成模型部署于云服务器的场景等。

[0195] 请参阅图17，其示出了本申请实施例提供的一种语音合成装置1700的结构框图。该装置1700从功能上划分，可以包括：信息获得模块1710、确定模块1720、模型获得模块1730以及语音合成模块1740。

[0196] 其中，信息获得模块1710获得针对语音合成模型的性能要求信息。

[0197] 确定模块1720用于根据所述性能要求信息，分别确定与所述性能要求信息对应的目标编码器、目标注意力组件和目标解码器。

[0198] 模型获得模块1730用于获得包括所述目标编码器、所述目标注意力组件和所述目标解码器的目标语音合成模型。

[0199] 可选地，模型获得模块1730具体可以用于：确定语音合成模型的目标框架，并按照目标框架对目标编码器、目标注意力组件和目标解码器进行组合，得到目标语音合成模型。

[0200] 语音合成模块1740用于通过所述目标语音合成模型将接收的字符信息合成为语音信息。

[0201] 可选地，确定模块1720具体可以用于：当性能要求信息包括第一性能指标信息，将具有残差网络结构的编码器确定为所述目标编码器，其中，所述残差网络结构包括依次连接的第一编码层和第二编码层，所述第一编码层的输出信息被叠加至所述第二编码层的输出信息。

[0202] 对应地，目标编码器处理字符信息的方式可以是：通过所述第一编码层，按照接收次序对所述字符信息中的每个字符进行编码，得到第一字符特征向量；通过所述第二编码层处理所述第一字符特征向量，得到第二字符特征向量；拼接所述第一字符特征向量和所

述第二字符特征向量，得到当前编码次序的字符对应的目标字符特征向量。

[0203] 可选地，确定模块1720具体还可以用于：当所述性能要求信息包括第二性能指标信息，将包括参数拟合层的注意力组件确定为所述目标注意力组件，其中，所述参数拟合层用于拟合高斯混合模型的参数特征信息，所述高斯混合模型是所述目标编码器在每个编码次序编码的字符与所述目标解码器在当前解码次序解码的字符的相关程度所服从的概率分布模型。

[0204] 对应地，目标注意力组件确定高斯混合模型的模型参数的方式可以是：通过所述参数拟合层处理所述目标编码器在各编码次序的隐藏状态以及所述目标解码器在当前解码次序的隐藏状态，得到所述高斯混合模型的第一参数特征信息、第二参数特征信息和第三参数特征信息；通过软最大化函数处理所述第一参数特征信息，得到所述高斯混合模型中每个高斯分布的权重；根据所述第二参数特征信息得到所述高斯混合模型中每个高斯分布的方差；通过软加函数处理所述第三参数特征信息，得到所述高斯混合模型中每个高斯分布的均值。

[0205] 进一步地，目标注意力组件处理目标编码器输出的目标字符特征向量的方式可以是：根据所述高斯混合模型中每个高斯分布的权重、方差和均值，得到所述高斯混合模型在当前解码次序的概率分布函数；根据所述概率分布函数和所述目标编码器在每个编码次序输出的目标字符特征向量，得到该目标字符特征向量在当前解码次序的注意力得分；根据所述目标编码器在各编码次序输出的目标字符向量及每个目标字符向量在当前解码次序的注意力得分，得到所述目标解码器在当前解码次序的注意力向量。

[0206] 可选地，确定模块1720具体还可以用于：当所述性能要求信息包括第三性能指标信息，将包括依次连接的循环门单元GRU层和长短时记忆网络LSTM层的解码器确定为所述目标解码器。

[0207] 对应地，目标解码器获得预测声学特征信息的方式可以是：确定当前解码次序的注意力向量；通过所述GRU层，对所述前一解码次序的注意力向量、所述当前解码次序的目标字符特征向量及所述目标解码器在前一解码次序的解码信息进行处理，得到第一声学特征向量；通过所述LSTM层处理所述第一声学特征及所述当前解码次序的注意力向量，得到第二声学特征向量；根据所述第二声学特征向量得到并输出所述目标解码器在所述当前解码次序的预测声学特征信息。

[0208] 可选地，语音合成装置1700还可以包括训练模块。

[0209] 训练模块可以用于在语音合成模块1740通过所述目标语音合成模型将接收的字符信息合成为语音信息之前，获得针对语音合成模型的音色要求信息；根据所述音色要求信息获取声音数据；基于所述声音数据对所述目标语音合成模型进行模型训练，使所述目标语音合成模型的第一损失函数达到优化条件。

[0210] 其中，第一损失函数可以通过如下方式建立：获取所述目标编码器在各编码次序输出的目标字符向量各自在当前解码次序的注意力得分，得到一注意力得分序列；确定所述注意力得分序列的熵；将所述熵叠加至第二损失函数，得到所述第一损失函数。

[0211] 可选地，将所述熵叠加至第二损失函数的方式可以是：将所述熵与目标权重的乘积叠加至所述第二损失函数。其中，所述目标权重在所述模型训练的过程中随迭代次数的增大而增大。

[0212] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述装置和模块的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0213] 在本申请所提供的几个实施例中,所显示或讨论的模块相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或模块的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0214] 另外,在本申请各个实施例中的各功能模块可以集成在一个处理模块中,也可以是各个模块单独物理存在,也可以两个或两个以上模块集成在一个模块中。上述集成的模块既可以采用硬件的形式实现,也可以采用软件功能模块的形式实现。

[0215] 请参考图18,其示出了本申请实施例提供的一种服务器200的结构框图。本申请中的服务器200可以包括一个或多个如下部件:处理器210、存储器220、以及一个或多个程序,其中一个或多个程序可以被存储在存储器220中并被配置为由一个或多个处理器210执行,一个或多个程序配置用于执行如前述方法实施例所描述的方法。

[0216] 处理器210可以包括一个或者多个处理核。处理器110利用各种接口和线路连接整个服务器200内的各个部分,通过运行或执行存储在存储器220内的指令、程序、代码集或指令集,以及调用存储在存储器220内的数据,执行服务器200的各种功能和处理数据。可选地,处理器210可以采用数字信号处理(Digital Signal Processing,DSP)、现场可编程门阵列(Field－Programmable Gate Array,FPGA)、可编程逻辑阵列(Programmable Logic Array,PLA)中的至少一种硬件形式来实现。处理器110可集成中央处理器(Central Processing Unit,CPU)、图像处理器(Graphics Processing Unit,GPU)和调制解调器等中的一种或几种的组合。其中,CPU主要处理操作系统、用户界面和应用程序等;GPU用于负责显示内容的渲染和绘制;调制解调器用于处理无线通信。可以理解的是,上述调制解调器也可以不集成到处理器210中,单独通过一块通信芯片进行实现。

[0217] 存储器220可以包括随机存储器(Random Access Memory,RAM),也可以包括只读存储器(Read-Only Memory)。存储器220可用于存储指令、程序、代码、代码集或指令集。存储器220可包括存储程序区和存储数据区,其中,存储程序区可存储用于实现操作系统的指令、用于实现至少一个功能的指令(比如触控功能、声音播放功能、图像播放功能等)、用于实现下述各个方法实施例的指令等。存储数据区还可以存储终端100在使用中所创建的数据(比如性能要求信息、目标语音合成模型)等。

[0218] 请参考图19,其示出了本申请实施例提供的一种计算机可读存储介质的结构框图。该计算机可读介质1900中存储有程序代码,所述程序代码可被处理器调用执行上述方法实施例中所描述的方法。

[0219] 计算机可读存储介质1900可以是诸如闪存、EEPROM(电可擦除可编程只读存储器)、EPROM、硬盘或者ROM之类的电子存储器。可选地,计算机可读存储介质1900包括非瞬时性计算机可读介质(non-transitory computer-readable storage medium)。计算机可读存储介质1900具有执行上述方法中的任何方法步骤的程序代码1910的存储空间。这些程序代码可以从一个或者多个计算机程序产品中读出或者写入到这一个或者多个计算机程序产品中。程序代码1910可以例如以适当形式进行压缩。

[0220] 最后应说明的是:以上实施例仅用以说明本申请的技术方案,而非对其限制;尽管参照前述实施例对本申请进行了详细的说明,本领域的普通技术人员当理解:其依然可以

对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不驱使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

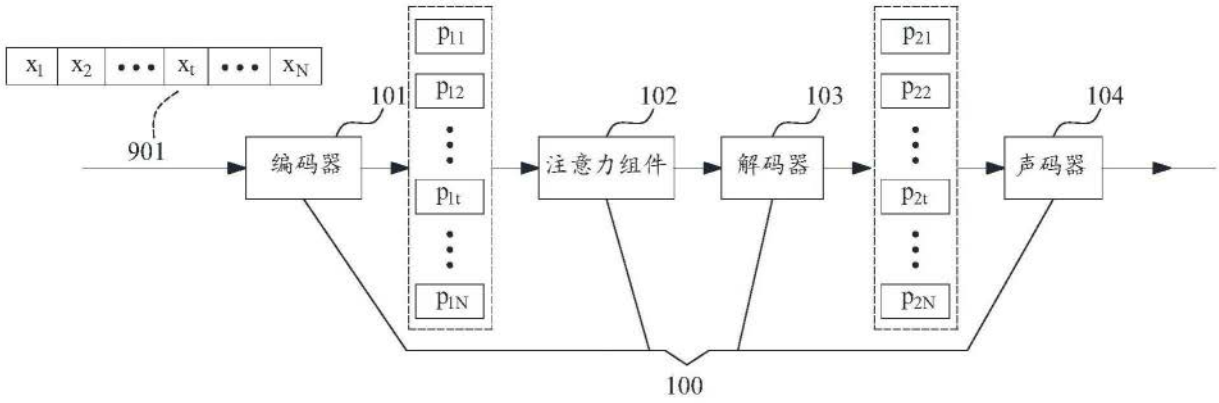

x1
x2
xt
xN
901
编码器
101
p11
p12
p1t
p1N
注意力组件
102
解码器
103
p21
p22
p2t
p2N
声码器
104
100

图1

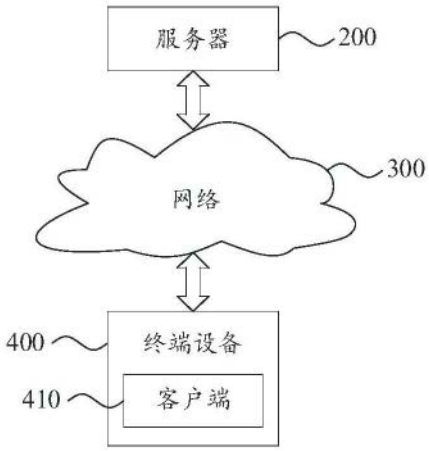

服务器
200
网络
300
400
终端设备
410
客户端

图2

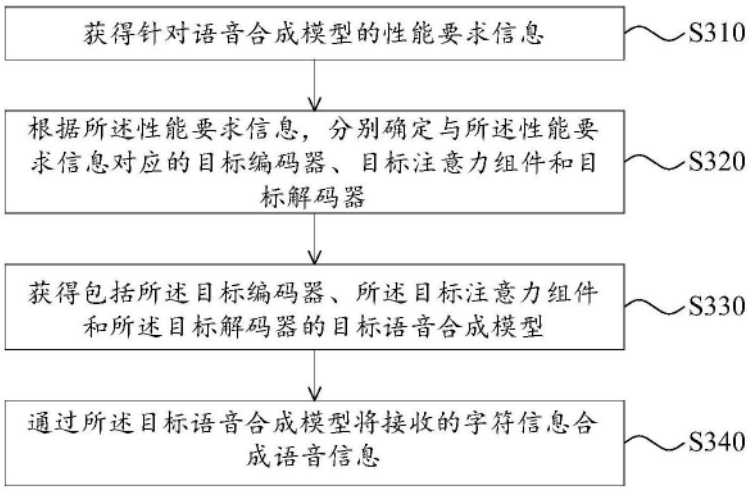

获得针对语音合成模型的性能要求信息
S310
根据所述性能要求信息，分别确定与所述性能要求信息对应的目标编码器、目标注意力组件和目标解码器
S320
获得包括所述目标编码器、所述目标注意力组件和所述目标解码器的目标语音合成模型
S330
通过所述目标语音合成模型将接收的字符信息合成语音信息
S340

图3

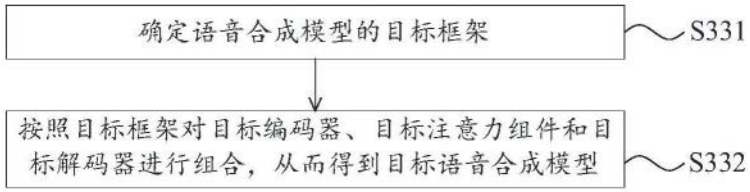

确定语音合成模型的目标框架
S331
按照目标框架对目标编码器、目标注意力组件和目标解码器进行组合，从而得到目标语音合成模型
S332

图4

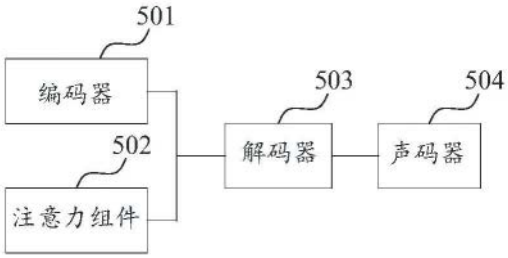

501
编码器
502
注意力组件
503
解码器
504
声码器

图5

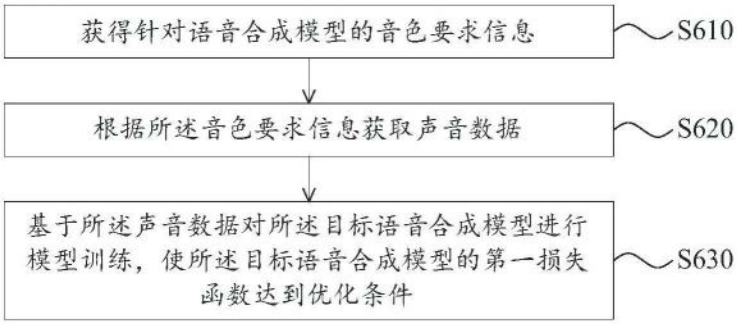

获得针对语音合成模型的音色要求信息
S610
根据所述音色要求信息获取声音数据
S620
基于所述声音数据对所述目标语音合成模型进行模型训练，使所述目标语音合成模型的第一损失函数达到优化条件
S630

图6

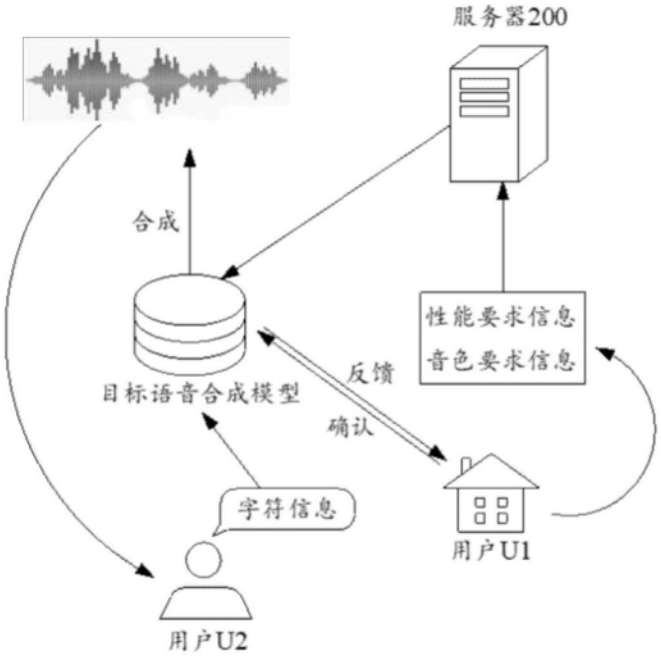

服务器200
合成
目标语音合成模型
反馈
确认
性能要求信息
音色要求信息
字符信息
用户U1
用户U2

图7

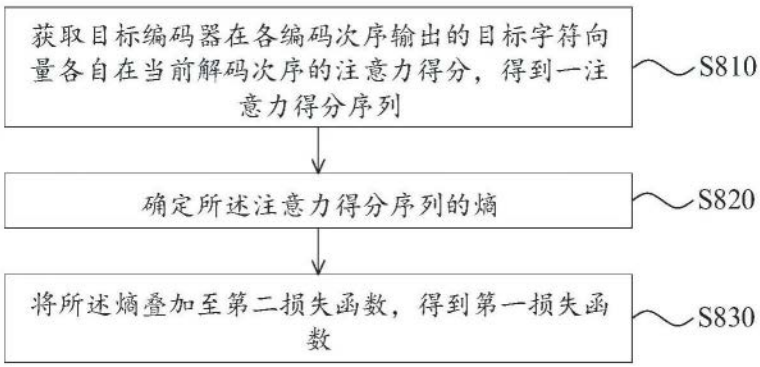

获取目标编码器在各编码次序输出的目标字符向量各自在当前解码次序的注意力得分，得到一注意力得分序列
S810
确定所述注意力得分序列的熵
S820
将所述熵叠加至第二损失函数，得到第一损失函数
S830

图8

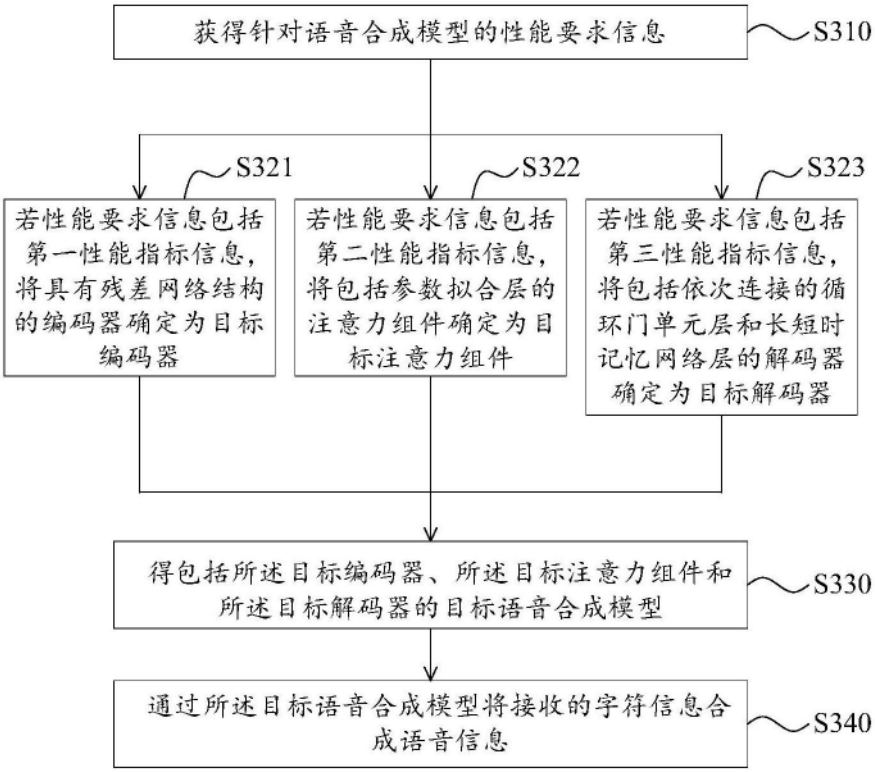

获得针对语音合成模型的性能要求信息
S310
S321
S322
S323
若性能要求信息包括第一性能指标信息，将具有残差网络结构的编码器确定为目标编码器
若性能要求信息包括第二性能指标信息，将包括参数拟合层的注意力组件确定为目标注意力组件
若性能要求信息包括第三性能指标信息，将包括依次连接的循环门单元层和长短时记忆网络层的解码器确定为目标解码器
得包括所述目标编码器、所述目标注意力组件和所述目标解码器的目标语音合成模型
S330
通过所述目标语音合成模型将接收的字符信息合成语音信息
S340

图9

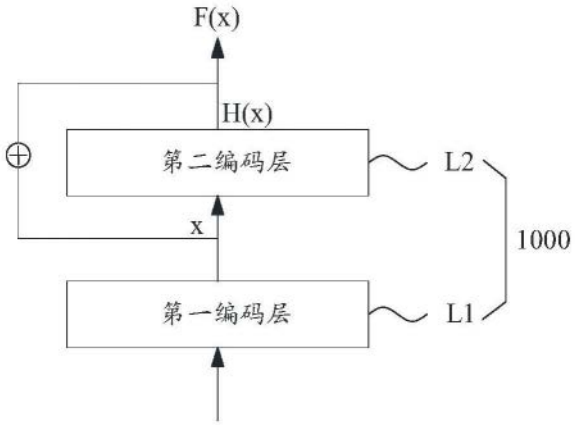

F(x)
H(x)
第二编码层
L2
x
1000
第一编码层
L1

图10A

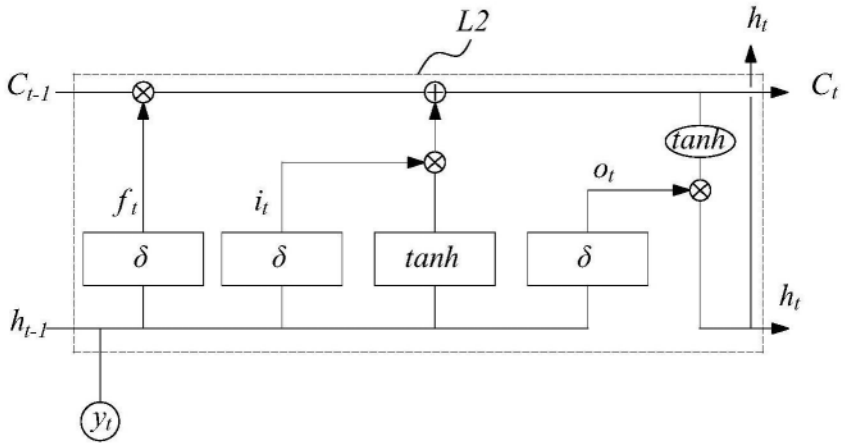

L2
h_t
C_{t-1}
C_t
tanh
o_t
f_t
i_t
δ
δ
tanh
δ
h_t
h_{t-1}
y_t

图10B

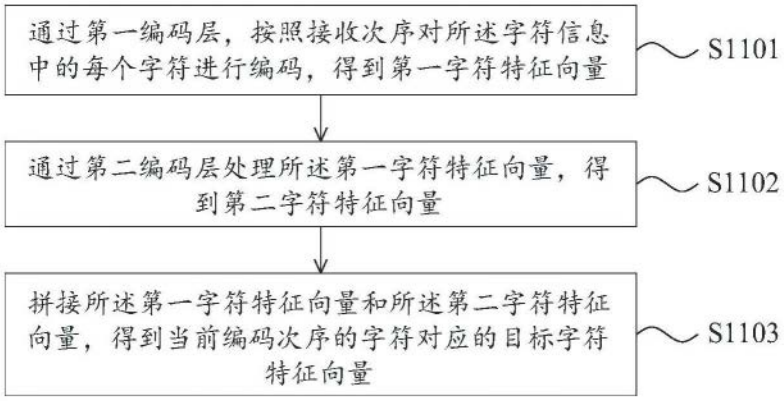

通过第一编码层，按照接收次序对所述字符信息中的每个字符进行编码，得到第一字符特征向量
S1101
通过第二编码层处理所述第一字符特征向量，得到第二字符特征向量
S1102
拼接所述第一字符特征向量和所述第二字符特征向量，得到当前编码次序的字符对应的目标字符特征向量
S1103

图11

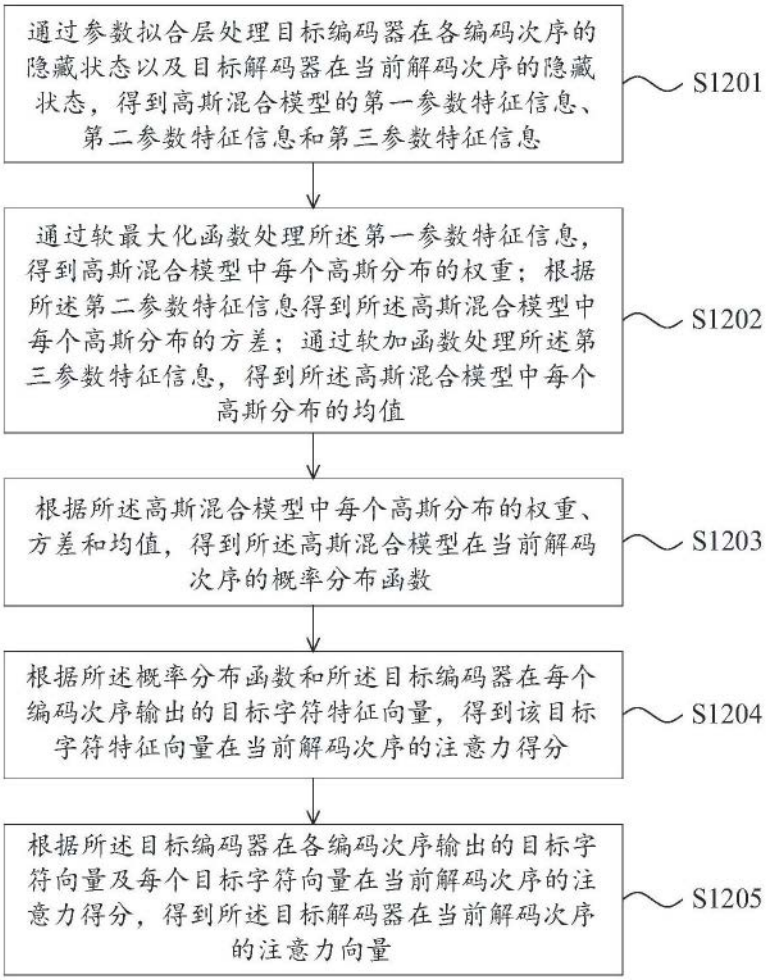

通过参数拟合层处理目标编码器在各编码次序的隐藏状态以及目标解码器在当前解码次序的隐藏状态，得到高斯混合模型的第一参数特征信息、第二参数特征信息和第三参数特征信息
S1201
通过软最大化函数处理所述第一参数特征信息，得到高斯混合模型中每个高斯分布的权重；根据所述第二参数特征信息得到所述高斯混合模型中每个高斯分布的方差；通过软加函数处理所述第三参数特征信息，得到所述高斯混合模型中每个高斯分布的均值
S1202
根据所述高斯混合模型中每个高斯分布的权重、方差和均值，得到所述高斯混合模型在当前解码次序的概率分布函数
S1203
根据所述概率分布函数和所述目标编码器在每个编码次序输出的目标字符特征向量，得到该目标字符特征向量在当前解码次序的注意力得分
S1204
根据所述目标编码器在各编码次序输出的目标字符向量及每个目标字符向量在当前解码次序的注意力得分，得到所述目标解码器在当前解码次序的注意力向量
S1205

图12

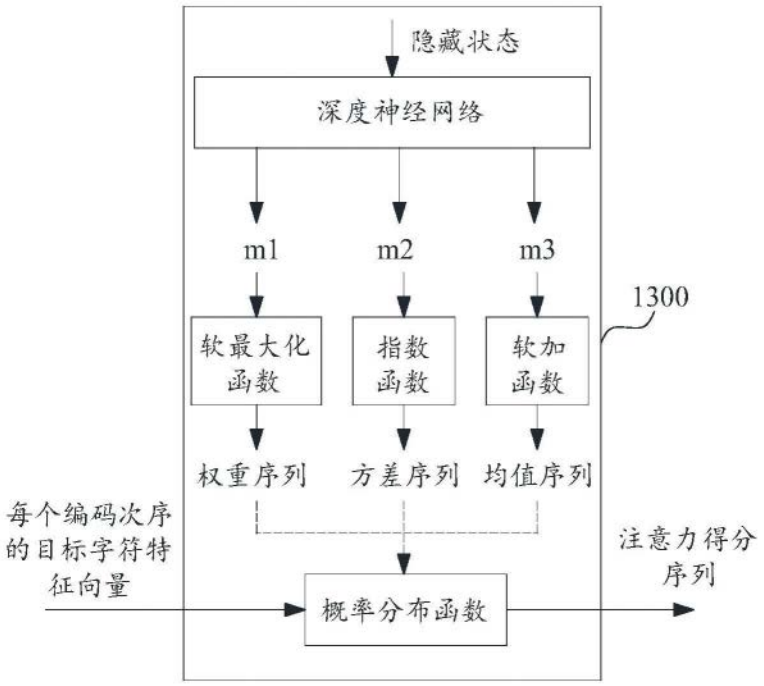

隐藏状态
深度神经网络
m1
m2
m3
1300
软最大化函数
指数函数
软加函数
权重序列
方差序列
均值序列
每个编码次序的目标字符特征向量
注意力得分序列
概率分布函数

图13

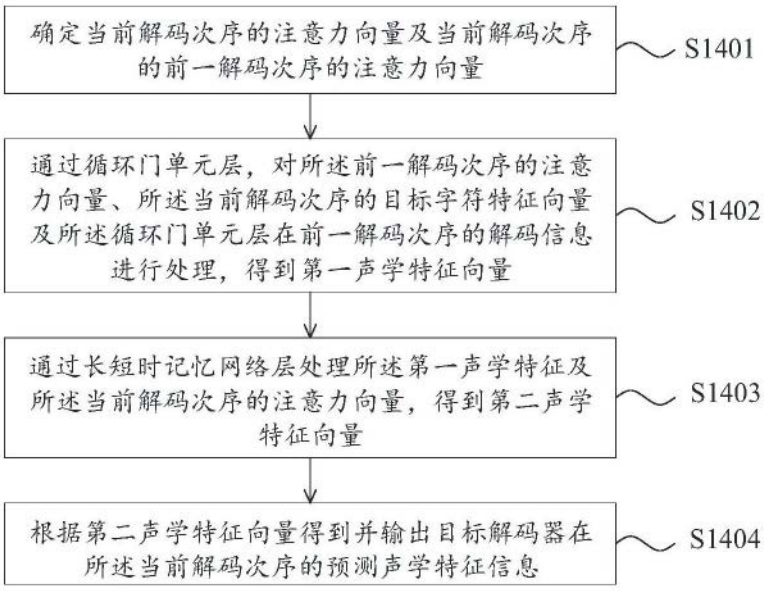

确定当前解码次序的注意力向量及当前解码次序的前一解码次序的注意力向量
S1401
通过循环门单元层，对所述前一解码次序的注意力向量、所述当前解码次序的目标字符特征向量及所述循环门单元层在前一解码次序的解码信息进行处理，得到第一声学特征向量
S1402
通过长短时记忆网络层处理所述第一声学特征及所述当前解码次序的注意力向量，得到第二声学特征向量
S1403
根据第二声学特征向量得到并输出目标解码器在所述当前解码次序的预测声学特征信息
S1404

图14

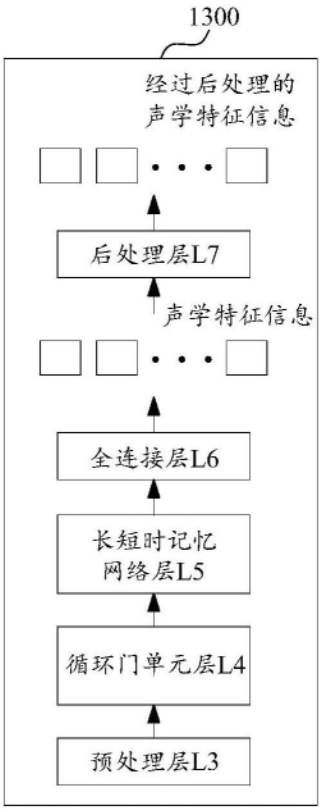

1300
经过后处理的
声学特征信息
后处理层L7
声学特征信息
全连接层L6
长短时记忆
网络层L5
循环门单元层L4
预处理层L3

图15A

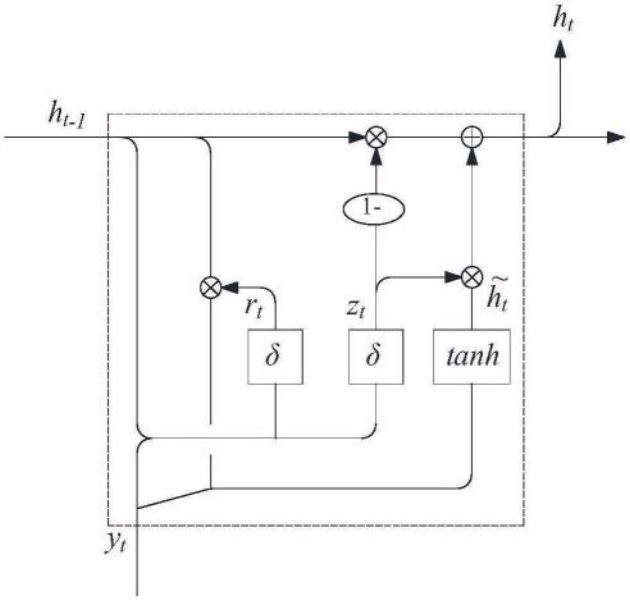

h_t
h_{t-1}
1-
r_t
z_t
$\tilde{h}_t$
δ
δ
tanh
y_t

图15B

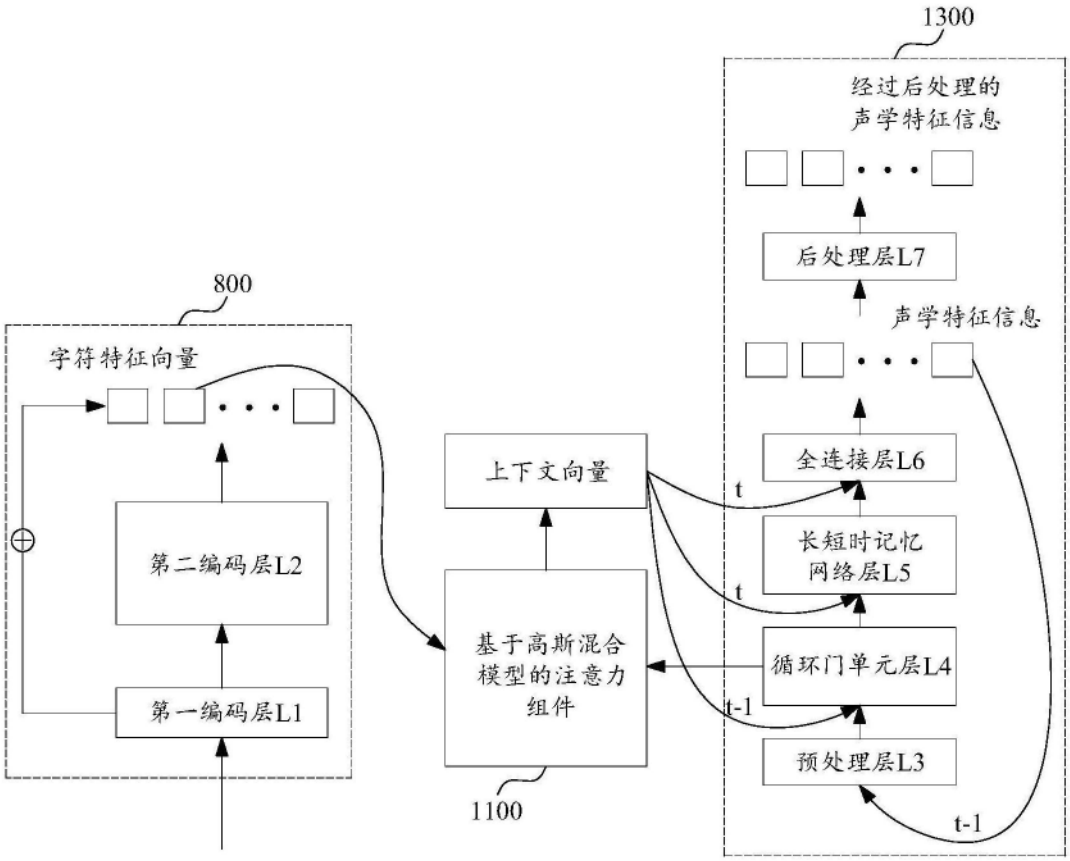

1300
经过后处理的
声学特征信息
后处理层L7
声学特征信息
全连接层L6
长短时记忆
网络层L5
循环门单元层L4
预处理层L3
t-1
800
字符特征向量
第二编码层L2
第一编码层L1
上下文向量
基于高斯混合
模型的注意力
组件
1100
t
t
t-1

图16

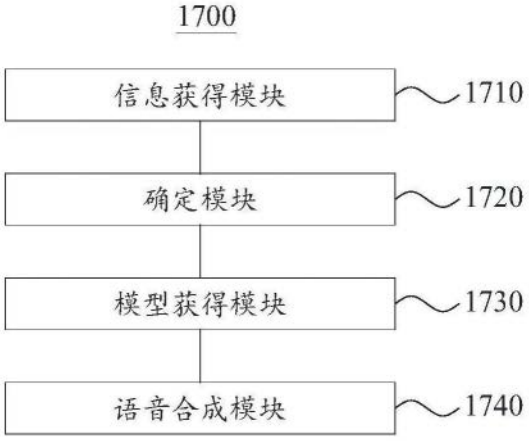

1700
信息获得模块
1710
确定模块
1720
模型获得模块
1730
语音合成模块
1740

图17

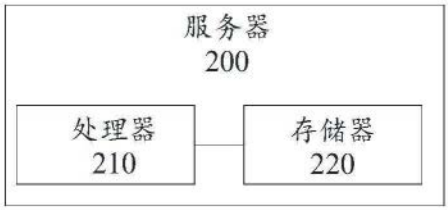

服务器
200
处理器
210
存储器
220

图18


计算机可读存储介质1900
1910
用于执行根据本申请方法实施
例各步骤的程序代码

图19