(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 112735454 A
(43)申请公布日 2021.04.30

(21)申请号 202011613263.1

(22)申请日 2020.12.30

(71)申请人 北京大米科技有限公司
地址 100142 北京市海淀区清河安宁庄东路18号23号楼二层2223

(72)发明人 梁光 舒景辰 吴雨璇 杨惠 周鼎皓

(74)专利代理机构 北京睿派知识产权代理事务所(普通合伙) 11597
代理人 刘锋

(51)Int.Cl.
G10L 21/02(2013.01)
G10L 21/0208(2013.01)
G10L 13/10(2013.01)
G10L 25/30(2013.01)
G10L 25/90(2013.01)

权利要求书2页 说明书9页 附图6页

(54)发明名称

音频处理方法、装置、电子设备和可读存储介质

(57)摘要

本发明实施例提供了一种音频处理方法、装置、电子设备和可读存储介质，涉及计算机技术领域。通过本发明实施例，基于原始音频样本训练的信息补偿模型具有较好的信息补偿能力，当使用训练后的信息补偿模型对待处理音频进行信息补偿时，可以使得目标音频中被补偿的部分与真实声音的相似度更高，进而使得目标音频的真实度更高，也就是说，训练后的信息补偿模型具有较高的升采样准确率。

CN 112735454 A

1.一种音频处理方法,其特征在于,所述方法包括:

获取待处理音频;以及

将所述待处理音频输入至预先训练的信息补偿模型进行处理,以获取目标音频;

其中,所述信息补偿模型基于如下步骤训练:

获取训练集,所述训练集包括多个样本组,所述样本组包括经降维处理后的第一音频样本和所述第一音频样本对应的原始音频样本;以及

根据所述训练集训练所述信息补偿模型。

2.根据权利要求1所述的方法,其特征在于,所述获取待处理音频,包括:

获取原始音频数据;以及

对所述原始音频数据进行降采样处理,获取待处理音频。

3.根据权利要求1所述的方法,其特征在于,所述第一音频样本中包括预设的噪声数据。

4.根据权利要求3所述的方法,其特征在于,所述噪声数据包括白噪声和/或粉红噪声。

5.根据权利要求3或4所述的方法,其特征在于,所述获取训练集,包括:

获取多个原始音频样本;

对于一原始音频样本,对所述原始音频样本进行降采样处理,获取第一音频数据;以及

将多个预设的噪声数据分别与所述第一音频数据进行组合,确定对应的多个第一音频样本,以获取所述原始音频样本对应的多个样本组。

6.根据权利要求1所述的方法,其特征在于,所述将所述待处理音频输入至预先训练的信息补偿模型进行处理,以获取目标音频,包括:

将所述待处理音频输入至预先训练的信息补偿模型进行升采样处理,以确定所述目标音频。

7.根据权利要求2所述的方法,其特征在于,所述获取原始音频数据,包括:

获取输入文本;

确定所述输入文本中至少一个字的发音向量,所述发音向量至少包括对应字的韵律信息;

确定各所述发音向量对应的发音时长以及发音音调,所述发音时长用于表征发音的持续时长,所述发音音调用于表征发音的音高;以及

基于所述发音向量、所述发音时长以及所述发音音调,合成所述输入文本对应的原始音频数据。

8.根据权利要求7所述的方法,其特征在于,所述发音音调为方言音调,所述方言音调用于表征方言发音的音高。

9.根据权利要求1所述的方法,其特征在于,所述信息补偿模型基于自回归神经网络或者生成对抗网络构建。

10.一种音频处理装置,其特征在于,所述装置包括:

第一获取模块,用于获取待处理音频;以及

信息补偿模块,用于将所述待处理音频输入至预先训练的信息补偿模型进行处理,以获取目标音频;

其中,所述信息补偿模型基于如下模块训练:

第二获取模块，用于获取训练集，所述训练集包括多个样本组，所述样本组包括经降维处理后的第一音频样本和所述第一音频样本对应的原始音频样本；以及

训练模块，用于根据所述训练集训练所述信息补偿模型。

11.根据权利要求10所述的装置，其特征在于，所述第一获取模块，具体用于：

获取原始音频数据；以及

对所述原始音频数据进行降采样处理，获取待处理音频。

12.根据权利要求10所述的装置，其特征在于，所述第一音频样本中包括预设的噪声数据。

13.根据权利要求12所述的装置，其特征在于，所述噪声数据包括白噪声和/或粉红噪声。

14.根据权利要求12或13所述的装置，其特征在于，所述第二获取模块，具体用于：

获取多个原始音频样本；

对于一原始音频样本，对所述原始音频样本进行降采样处理，获取第一音频数据；以及

将多个预设的噪声数据分别与所述第一音频数据进行组合，确定对应的多个第一音频样本，以获取所述原始音频样本对应的多个样本组。

15.根据权利要求10所述的装置，其特征在于，所述信息补偿模块，具体用于：

将所述待处理音频输入至预先训练的信息补偿模型进行升采样处理，以确定所述目标音频。

16.根据权利要求11所述的装置，其特征在于，所述第一获取模块，具体还用于：

获取输入文本；

确定所述输入文本中至少一个字的发音向量，所述发音向量至少包括对应字的韵律信息；

确定各所述发音向量对应的发音时长以及发音音调，所述发音时长用于表征发音的持续时长，所述发音音调用于表征发音的音高；以及

基于所述发音向量、所述发音时长以及所述发音音调，合成所述输入文本对应的原始音频数据。

17.根据权利要求16所述的装置，其特征在于，所述发音音调为方言音调，所述方言音调用于表征方言发音的音高。

18.根据权利要求10所述的装置，其特征在于，所述信息补偿模型基于自回归神经网络或者生成对抗网络构建。

19.一种电子设备，包括存储器和处理器，其特征在于，所述存储器用于存储一条或多条计算机程序指令，其中，所述一条或多条计算机程序指令被所述处理器执行以实现如权利要求1-9中任一项所述的方法。

20.一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内存储有计算机程序，所述计算机程序被处理器执行时实现权利要求1-9任一项所述的方法。

音频处理方法、装置、电子设备和可读存储介质

技术领域

[0001] 本申请涉及计算机技术领域，特别是涉及一种音频处理方法、装置、电子设备和可读存储介质。

背景技术

[0002] 目前，音频处理可以应用在各种场景，例如对机器合成语音(在线教育、视频配音以及解说等等)进行音频处理，在实际应用过程中，常见的音频处理包括音频数据压缩以及音频数据还原。

[0003] 然而，在音频数据压缩以及音频数据还原的过程中，往往会对音频数据产生数据损耗，降低了音频数据还原的准确率。

发明内容

[0004] 有鉴于此，本发明实施例提供一种音频处理方法、装置、电子设备和可读存储介质，以使得信息补偿模型具有较好的信息补偿能力和较高的升采样准确率。

[0005] 第一方面，提供了一种音频处理方法，所述方法应用于电子设备，所述方法包括：

[0006] 获取待处理音频。

[0007] 将所述待处理音频输入至预先训练的信息补偿模型进行处理，以获取目标音频。

[0008] 其中，所述信息补偿模型基于如下步骤训练：

[0009] 获取训练集，所述训练集包括多个样本组，所述样本组包括经降维处理后的第一音频样本和所述第一音频样本对应的原始音频样本。

[0010] 根据所述训练集训练所述信息补偿模型。

[0011] 可选的，所述获取待处理音频，包括：

[0012] 获取原始音频数据。

[0013] 对所述原始音频数据进行降采样处理，获取待处理音频。

[0014] 可选的，所述第一音频样本中包括预设的噪声数据。

[0015] 可选的，所述噪声数据包括白噪声和/或粉红噪声。

[0016] 可选的，所述获取训练集，包括：

[0017] 获取多个原始音频样本。

[0018] 对于一原始音频样本，对所述原始音频样本进行降采样处理，获取第一音频数据。

[0019] 将多个预设的噪声数据分别与所述第一音频数据进行组合，确定对应的多个第一音频样本，以获取所述原始音频样本对应的多个样本组。

[0020] 可选的，所述将所述待处理音频输入至预先训练的信息补偿模型进行处理，以获取目标音频，包括：

[0021] 将所述待处理音频输入至预先训练的信息补偿模型进行升采样处理，以确定所述目标音频。

[0022] 可选的，所述获取原始音频数据，包括：

[0023] 获取输入文本。

[0024] 确定所述输入文本中至少一个字的发音向量，所述发音向量至少包括对应字的韵律信息。

[0025] 确定各所述发音向量对应的发音时长以及发音音调，所述发音时长用于表征发音的持续时长，所述发音音调用于表征发音的音高。

[0026] 基于所述发音向量、所述发音时长以及所述发音音调，合成所述输入文本对应的原始音频数据。

[0027] 可选的，所述发音音调为方言音调，所述方言音调用于表征方言发音的音高。

[0028] 可选的，所述信息补偿模型基于自回归神经网络或者生成对抗网络构建。

[0029] 第二方面，提供了一种音频处理装置，所述装置应用于电子设备，所述装置包括：

[0030] 第一获取模块，用于获取待处理音频。

[0031] 信息补偿模块，用于将所述待处理音频输入至预先训练的信息补偿模型进行处理，以获取目标音频。

[0032] 其中，所述信息补偿模型基于如下模块训练：

[0033] 第二获取模块，用于获取训练集，所述训练集包括多个样本组，所述样本组包括经降维处理后的第一音频样本和所述第一音频样本对应的原始音频样本。

[0034] 训练模块，用于根据所述训练集训练所述信息补偿模型。

[0035] 可选的，所述第一获取模块，具体用于：

[0036] 获取原始音频数据。

[0037] 对所述原始音频数据进行降采样处理，获取待处理音频。

[0038] 可选的，所述第一音频样本中包括预设的噪声数据。

[0039] 可选的，所述噪声数据包括白噪声和/或粉红噪声。

[0040] 可选的，所述第二获取模块，具体用于：

[0041] 获取多个原始音频样本。

[0042] 对于一原始音频样本，对所述原始音频样本进行降采样处理，获取第一音频数据。

[0043] 将多个预设的噪声数据分别与所述第一音频数据进行组合，确定对应的多个第一音频样本，以获取所述原始音频样本对应的多个样本组。

[0044] 可选的，所述信息补偿模块，具体用于：

[0045] 将所述待处理音频输入至预先训练的信息补偿模型进行升采样处理，以确定所述目标音频。

[0046] 可选的，所述第一获取模块，具体还用于：

[0047] 获取输入文本。

[0048] 确定所述输入文本中至少一个字的发音向量，所述发音向量至少包括对应字的韵律信息。

[0049] 确定各所述发音向量对应的发音时长以及发音音调，所述发音时长用于表征发音的持续时长，所述发音音调用于表征发音的音高。

[0050] 基于所述发音向量、所述发音时长以及所述发音音调，合成所述输入文本对应的原始音频数据。

[0051] 可选的，所述发音音调为方言音调，所述方言音调用于表征方言发音的音高。

[0052] 可选的，所述信息补偿模型基于自回归神经网络或者生成对抗网络构建。

[0053] 第三方面，本发明实施例提供了一种电子设备，包括存储器和处理器，所述存储器用于存储一条或多条计算机程序指令，其中，所述一条或多条计算机程序指令被所述处理器执行以实现如第一方面所述的方法。

[0054] 第四方面，本发明实施例提供了一种计算机可读存储介质，其上存储计算机程序指令，所述计算机程序指令在被处理器执行时实现如第一方面所述的方法。

[0055] 通过本发明实施例，基于原始音频样本训练的信息补偿模型具有较好的信息补偿能力，当使用训练后的信息补偿模型对待处理音频进行信息补偿时，可以使得目标音频中被补偿的部分与真实声音的相似度更高，进而使得目标音频的真实度更高，也就是说，训练后的信息补偿模型具有较高的升采样准确率。

附图说明

[0056] 通过以下参照附图对本发明实施例的描述，本发明实施例的上述以及其它目的、特征和优点将更为清楚，在附图中：

[0057] 图1为本发明实施例提供的一种相关技术中降维处理过程的示意图；

[0058] 图2为本发明实施例提供的一种音频处理方法的流程图；

[0059] 图3为本发明实施例提供的另一种音频处理方法的流程图；

[0060] 图4为本发明实施例提供的一种音频处理方法的示意图；

[0061] 图5为本发明实施例提供的一种确定第一音频样本过程的示意图；

[0062] 图6为本发明实施例提供的另一种音频处理方法的流程图；

[0063] 图7为本发明实施例提供的一种音频处理装置的结构示意图；

[0064] 图8为本发明实施例提供的另一种音频处理装置的结构示意图；

[0065] 图9为本发明实施例提供的一种电子设备的结构示意图。

具体实施方式

[0066] 以下基于实施例对本发明进行描述，但是本发明并不仅仅限于这些实施例。在下文对本发明的细节描述中，详尽描述了一些特定的细节部分。对本领域技术人员来说没有这些细节部分的描述也可以完全理解本发明。为了避免混淆本发明的实质，公知的方法、过程、流程、元件和电路并没有详细叙述。

[0067] 此外，本领域普通技术人员应当理解，在此提供的附图都是为了说明的目的，并且附图不一定是按比例绘制的。

[0068] 除非上下文明确要求，否则在说明书的“包括”、“包含”等类似词语应当解释为包含的含义而不是排他或穷举的含义；也就是说，是“包括但不限于”的含义。

[0069] 在本发明的描述中，需要理解的是，术语“第一”、“第二”等仅用于描述目的，而不能理解为指示或暗示相对重要性。此外，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。

[0070] 目前，音频处理可以应用在各种场景，例如对机器合成语音(在线教育、视频配音以及解说等等)进行音频处理，在得到原始音频数据后，为了减轻音频处理过程中的数据计算量，往往会对原始音频数据先进行降采样处理，然后在进行后续的处理，进而得到可以播

放的目标音频数据。

[0071] 在此过程中，对原始音频数据先进行降采样处理时，会对原始音频数据进行降维处理，进而降低该原始音频数据的数据量。

[0072] 例如，如图1所示，图1为本发明实施例提供的一种相关技术中降维处理过程的示意图，该示意图包括：原始音频11、中间音频12、待播发音频13和播放设备14。

[0073] 具体的，当电子设备获取到原始音频11后，可以对原始音频11进行降采样处理，以确定中间音频12。

[0074] 其中，电子设备可以是终端设备或者服务器，终端设备可以是智能手机、平板电脑或者个人计算机(Personal Computer，PC)等，服务器可以是单个服务器，也可以是以分布式方式配置的服务器集群，还可以是云服务器。

[0075] 在此过程中，为了保证音频处理过程的顺利进行，需要将原始音频11的维度降低至一个较低的数值，例如，原始音频11是采样率为22K*16的音频数据，为了使得电子设备可以更有效的对原始音频11进行音频处理，需要对原始音频11进行降采样处理，确定中间音频12(中间音频12可以是80维的梅尔频谱，即中间音频12可以是80*1000的梅尔频谱)。

[0076] 如图1所示，当原始音频11经过降采样处理得到中间音频12后，由于原始音频11经过数据压缩，所以，中间音频12与原始音频11相比会丢失一部分数据(例如图1中的中间音频12丢失了原始音频11中的高频部分)。

[0077] 当确定中间音频12后，电子设备可以将中间音频12输入声码器，声码器在接收到中间音频12后，可以对中间音频12进行升采样处理，得到待播发音频13，进而，电子设备可以通过播放设备14播放待播发音频13。

[0078] 其中，播放设备14既可以是安装于电子设备的音频播放设备，也可以是外接于电子设备的音频播放设备，本发明实施例对此不做限定。

[0079] 在声码器对中间音频12进行升采样处理时，由于中间音频12相较于原始音频11丢失了大量数据，所以声码器升采样处理后得到的待播发音频13与原始音频11相比有较大差别，即还原音频数据的准确率较低。

[0080] 为了提高提高音频数据处理过程中升采样的准确率，本发明实施例提供一种音频处理方法，该方法应用于电子设备，如图2所示，该方法包括如下步骤：

[0081] 在步骤21，获取待处理音频。

[0082] 在本发明实施例中，待处理音频既可以是经过降采样处理后得到的音频数据，也可以是未经过降采样处理的音频数据。

[0083] 其中，若待处理音频既是经过降采样处理后得到的音频数据，则该降采样处理的过程可以执行为：获取原始音频数据，以及对原始音频数据进行降采样处理，获取待处理音频。

[0084] 在实际应用中，可以通过特定工具对原始音频数据进行降采样处理，例如，可以通过FFmpeg(Fast Forward Mpeg)对原始音频进行降采样处理，其中，FFmpeg是一套可以用来记录、转换数字音频、视频，并能将其转化为流的开源计算机程序，基于FFmpeg的功能，可以实现对原始音频数据的降采样，当然，也可以通过其他适用的工具、算法、模型等实现降采样处理，本发明实施例对此不做限定。

[0085] 通过本发明实施例，由于待处理音频既可以是经过降采样处理后得到的音频数

据,也可以是未经过降采样处理的音频数据,所以使得本发明实施例的适用性更强,也就是说,一段音频数据无论是否经过降采样处理都可以通过本发明实施例进行信息补充。

[0086] 在步骤22,将待处理音频输入至预先训练的信息补偿模型进行处理,以获取目标音频。

[0087] 其中,信息补偿模型可以对待处理音频进行数据补偿,补偿后的待处理音频的数据维度将得到提升,也就是说,在一种可选的实施方式中,步骤22可以执行为:将待处理音频输入至预先训练的信息补偿模型进行升采样处理,以确定目标音频。

[0088] 为了保证步骤22中的信息补偿模型的补偿能力,在本发明实施例中需要对该信息补偿模型进行训练,如图3所示,信息补偿模型基于如下步骤训练:

[0089] 在步骤31,获取训练集。

[0090] 其中,该训练集包括多个样本组,该样本组包括经降维处理后的第一音频样本和第一音频样本对应的原始音频样本。

[0091] 在本发明实施例中,因为信息补偿模型需要对音频数据准确的进行补偿,所以,在训练的过程中需要获取准确的正样本,进而,在本发明实施例中,将原始音频样本作为正样本可以使得信息补偿模型的训练得到良好的监督,最终使得信息补偿模型可以准确的进行信息补偿。

[0092] 在步骤32,根据该训练集训练信息补偿模型。

[0093] 通过本发明实施例,基于原始音频样本训练的信息补偿模型具有较好的信息补偿能力,当使用训练后的信息补偿模型对待处理音频进行信息补偿时,可以使得目标音频中被补偿的部分与真实声音的相似度更高,进而使得目标音频的真实度更高,也就是说,训练后的信息补偿模型具有较高的升采样准确率。

[0094] 为了更好的进行解释说明,本发明实施例提供一种上述音频处理方法过程的示意图,如图4所示,该示意图包括:待处理音频41、目标音频42、信息补偿模型43和损失函数44。

[0095] 由图4可知,在待处理音频41中缺失了高频部分的音频数据,其中,该待处理音频41既可以是经过降采样处理后得到的音频数据,也可以是未经过降采样处理的音频数据。

[0096] 当电子设备获取待处理音频41之后,可以基于信息补偿模型43对待处理音频41进行信息补偿,以确定目标音频42,具体的,在图4中,信息补偿模型43可以基于待处理音频41的低频部分,生成包括低频部分和高频部分的目标音频42,当然,在其他适用的情况下,信息补偿模型43也可以基于待处理音频的高频部分生成包括低频部分和高频部分的目标音频,本发明实施例对此不做限定。

[0097] 也就是说,在本发明实施例中,信息补偿模型可以是一种音频数据的生成模型,具体的,信息补偿模型可以基于自回归神经网络(WaveNet)或者生成对抗网络(melgan)构建。

[0098] 其中,WaveNet是一种概率自回归模型,其可以基于之前已经生成的所有样本,来预测当前音频样本的概率分布,也就是说,WaveNet可以基于待处理音频来预测目标音频的概率分布。

[0099] melgan是一种基于gan网络的生成对抗模型,其包括生成器和判别器两部分,其中,生成器可以用于生成音频数据,而判别器可以在训练的过程中用于判别melgan生成的音频数据是否是真实数据,然后,根据判断结果调整模型参数。

[0100] 另外,在图4中,信息补偿模型43进行训练时(即图4中虚线箭头所指的数据流向),

可以通过损失函数44计算目标音频42和正样本之间的损失，进而根据该损失调整信息补偿模型43，其中，损失函数44可以是交叉熵函数、反向传播算法等，本发明实施例对损失函数44不做限定。

[0101] 通过上述步骤31至步骤32对信息补偿模型的训练，可以使得信息补偿模型具有较好的信息补偿能力，在实际的音频处理过程中，不可避免的会存在一些噪声干扰，相关技术中，常用的声码器的抗噪能力较差，即相关技术中经过声码器升采样处理后的音频数据质量较差。

[0102] 为了提高本发明实施例中信息补偿模型的抗噪能力，可以在训练样本中加入预先设置的噪声数据，以提升信息补偿模型的抗噪能力。

[0103] 在一种可实施方式中，上述训练集中的第一音频样本可以包括预设的噪声数据，其中，该噪声数据可以是白噪声和/或粉红噪声，也可以是其它噪声，本发明实施例对此不做赘述。

[0104] 白噪声(white noise)是指功率谱密度在整个频域内是常数的噪声。粉红噪声是指每个倍频程的强度相等的噪声，即粉红噪声在一定的范围(倍频程)内具有相同或类似的能量。

[0105] 具体的，在一种可实施方式中，上述步骤31可以执行为：获取多个原始音频样本，对于一原始音频样本，对该原始音频样本进行降采样处理，获取第一音频数据，以及将多个预设的噪声数据分别与第一音频数据进行组合，确定对应的多个第一音频样本，以获取原始音频样本对应的多个样本组。

[0106] 例如，如图5所示，图5为本发明实施例提供的一种确定第一音频样本过程的示意图，该示意图包括：噪声数据A、噪声数据B、噪声数据C、第一音频数据51和多个第一音频样本52。

[0107] 当电子设备获取到原始音频样本后，可以通过降采样处理确定该原始音频对应的第一音频数据51，针对第一音频数据51，电子设备可以将预先设置各噪声数据与第一音频数据51进行组合，以确定多个第一音频样本52。

[0108] 具体的，如图所示，当电子设备确定第一音频数据51后，可以将预先设置的噪声数据A、噪声数据B和噪声数据C与第一音频数据51进行组合，以确定多个第一音频样本52，其中，第一音频数据51既可以和一个噪声数据进行组合，也可以和多个噪声数据进行组合，也就是说，在图5中的第一音频数据51和至少一个噪声数据进行组合时，可以确定7个第一音频样本52，具体包括：51+A+B+C、51+A+B、51+B+C、51+A+C、51+A、51+B和51+C，其中，“51”用于表征第一音频数据51，“A”用于表征噪声数据A，“B”用于表征噪声数据B，“C”用于表征噪声数据C。

[0109] 当确定多个第一音频样本52后，电子设备可以基于多个第一音频样本52确定原始音频样本对应的多个样本组，进而根据该多个样本组对信息补偿模型进行训练。

[0110] 通过本发明实施例，由于在训练样本中加入了预先设置的噪声数据，所以在训练的过程中可以使得信息补偿模型具有较好的去除噪声能力，当该信息补偿模型训练完成后，该信息补偿模型可以有效去除待处理音频中的噪声，进而使得目标音频具有更高的音质，也就是说，通过本发明实施例，可以使得训练后的信息补偿模型具有较好的抗噪能力。

[0111] 在本发明实施例中，训练后的信息补偿模型可以对待处理音频进行高质量的信息

补偿以获取目标音频，其中，待处理音频既可以是直接获取的音频数据，也可以是对原始音频数据进行降采样处理后得到的音频数据。

[0112] 其中，原始音频数据可以是通过电子设备进行语音合成得到的原始音频数据，具体的，如图6所示，原始音频数据可以通过如下步骤确定：

[0113] 在步骤61，获取输入文本。

[0114] 在实际应用中，语音合成可以针对至少一段输入文本中的文字内容确定每个文字对应的发音，进而根据每个文字的发音合成一段连续的语音。其中，输入文本中至少包括一个字。

[0115] 在步骤62，确定输入文本中至少一个字的发音向量。

[0116] 其中，发音向量至少包括对应字的韵律信息，发音向量可以用于表征输入文本中至少一个字的嵌入(embedding)，韵律信息可以用于表征对应字之后的停顿时长，其中，embedding是深度学习中常用的特征提取手段，具体的，特征提取就是把高维原始数据(图像、文字等)映射到低维流形(Manifold)，使得高维原始数据被映射到低维流形之后变得可分离，这个映射的过程可以叫做embedding，例如Word embedding，就是把单词组成的句子映射到一个表征向量，而在本发明实施例中，embedding的对象是输入文本中的字。

[0117] 在一种可实施方式中，步骤62可以执行为：基于预先设置的文字和拼音的对应关系，确定输入文本中至少一个字的拼音信息，以及对拼音信息进行向量化处理，确定拼音信息的发音向量。

[0118] 具体的，在本发明实施例中，可以基于字典等工具预先设置文字和拼音的对应关系，当接收到输入文本后，可以针对输入文本中的每个字，确定每个字对应的拼音，然后针对每个字的拼音分别进行Embedding处理，确定每个拼音的特征向量，然后将该特征向量作为对应字的发音向量。

[0119] 在步骤63，确定各发音向量对应的发音时长以及发音音调。

[0120] 其中，发音时长用于表征发音的持续时长，发音音调用于表征发音的音高。

[0121] 在本发明实施例中，发音时长可以基于带有长度调节器(Length Regulator)的发音时长预测模型预测得到，其中，长度调节器可以用于解决音素和频谱图序列之间的长度不匹配问题，基于长度调节器，可以使得模型能够准确预测每个音素所对应的持续时长。

[0122] 发音音调可以基于带有音高预测器(pitch predictor)的发音音调预测模型预测得到，其中，音高预测器可以基于卷积网络的卷积运算以及全连接层确定每个发音向量所对应的音高。另外，若发音音调预测模型用于预测发音向量的方言音调，则该发音音调预测模型中的音高预测器所输出的音高为每个发音向量所对应的方言音高。

[0123] 也就是说，在一种优选的实施方式中，上述发音音调可以是方言音调，方言音调用于表征方言发音的音高。在原始音频数据合成的过程中，将方言音调作为发音音调可以为原始音频数据附加方言独有的音高(也即方言独有的发音方式)，使得原始音频数据具有更加贴近人类的说话方式。

[0124] 在步骤64，基于发音向量、发音时长以及发音音调，合成输入文本对应的原始音频数据。

[0125] 通过本发明实施例，可以通过发音向量、韵律标签以及发音向量对应的发音时长，使得原始音频数据可以具有停顿、延长音等人类常用的说话形式，然后，还可以基于发音音

调为原始音频数据附加音高，使得原始音频数据更加贴近人类的说话方式，最终，基于发音向量、韵律标签、发音时长以及发音音调确定的原始音频数据可以与人声具有较高的相似度。

[0126] 基于相同的技术构思，本发明实施例还提供了一种音频处理装置，如图7所示，该装置包括：第一获取模块和信息补偿模块。

[0127] 第一获取模块71，用于获取待处理音频。

[0128] 信息补偿模块72，用于将所述待处理音频输入至预先训练的信息补偿模型进行处理，以获取目标音频。

[0129] 其中，如图8所示，所述信息补偿模型基于如下模块训练：第二获取模块81和训练模块82。

[0130] 第二获取模块81，用于获取训练集，所述训练集包括多个样本组，所述样本组包括经降维处理后的第一音频样本和所述第一音频样本对应的原始音频样本。

[0131] 训练模块82，用于根据所述训练集训练所述信息补偿模型。

[0132] 可选的，所述第一获取模块71，具体用于：

[0133] 获取原始音频数据。

[0134] 对所述原始音频数据进行降采样处理，获取待处理音频。

[0135] 可选的，所述第一音频样本中包括预设的噪声数据。

[0136] 可选的，所述噪声数据包括白噪声和/或粉红噪声。

[0137] 可选的，所述第二获取模块81，具体用于：

[0138] 获取多个原始音频样本。

[0139] 对于一原始音频样本，对所述原始音频样本进行降采样处理，获取第一音频数据。

[0140] 将多个预设的噪声数据分别与所述第一音频数据进行组合，确定对应的多个第一音频样本，以获取所述原始音频样本对应的多个样本组。

[0141] 可选的，所述信息补偿模块72，具体用于：

[0142] 将所述待处理音频输入至预先训练的信息补偿模型进行升采样处理，以确定所述目标音频。

[0143] 可选的，所述第一获取模块71，具体还用于：

[0144] 获取输入文本。

[0145] 确定所述输入文本中至少一个字的发音向量，所述发音向量至少包括对应字的韵律信息。

[0146] 确定各所述发音向量对应的发音时长以及发音音调，所述发音时长用于表征发音的持续时长，所述发音音调用于表征发音的音高。

[0147] 基于所述发音向量、所述发音时长以及所述发音音调，合成所述输入文本对应的原始音频数据。

[0148] 可选的，所述发音音调为方言音调，所述方言音调用于表征方言发音的音高。

[0149] 可选的，所述信息补偿模型基于自回归神经网络或者生成对抗网络构建。

[0150] 通过本发明实施例，基于原始音频样本训练的信息补偿模型具有较好的信息补偿能力，当使用训练后的信息补偿模型对待处理音频进行信息补偿时，可以使得目标音频中被补偿的部分与真实声音的相似度更高，进而使得目标音频的真实度更高，也就是说，训练

后的信息补偿模型具有较高的升采样准确率。

[0151] 图9是本发明实施例的电子设备的示意图。如图9所示，图9所示的电子设备为通用地址查询装置，其包括通用的计算机硬件结构，其至少包括处理器91和存储器92。处理器91和存储器92通过总线93连接。存储器92适于存储处理器91可执行的指令或程序。处理器91可以是独立的微处理器，也可以是一个或者多个微处理器集合。由此，处理器91通过执行存储器92所存储的指令，从而执行如上所述的本发明实施例的方法流程实现对于数据的处理和对于其它装置的控制。总线93将上述多个组件连接在一起，同时将上述组件连接到显示控制器94和显示装置以及输入/输出(I/O)装置95。输入/输出(I/O)装置95可以是鼠标、键盘、调制解调器、网络接口、触控输入装置、体感输入装置、打印机以及本领域公知的其他装置。典型地，输入/输出装置95通过输入/输出(I/O)控制器96与系统相连。

[0152] 本领域的技术人员应明白，本发明的实施例可提供为方法、装置(设备)或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可读存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品。

[0153] 本发明是参照根据本发明实施例的方法、装置(设备)和计算机程序产品的流程图来描述的。应理解可由计算机程序指令实现流程图中的每一流程。

[0154] 这些计算机程序指令可以存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现流程图一个流程或多个流程中指定的功能。

[0155] 也可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程中指定的功能的装置。

[0156] 本发明的另一实施例涉及一种非易失性存储介质，用于存储计算机可读程序，所述计算机可读程序用于供计算机执行上述部分或全部的方法实施例。

[0157] 即，本领域技术人员可以理解，实现上述实施例方法中的全部或部分步骤是可以通过程序来指定相关的硬件来完成，该程序存储在一个存储介质中，包括若干指令用以使得一个设备(可以是单片机，芯片等)或处理器(processor)执行本发明各实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0158] 以上所述仅为本发明的优选实施例，并不用于限制本发明，对于本领域技术人员而言，本发明可以有各种改动和变化。凡在本发明的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

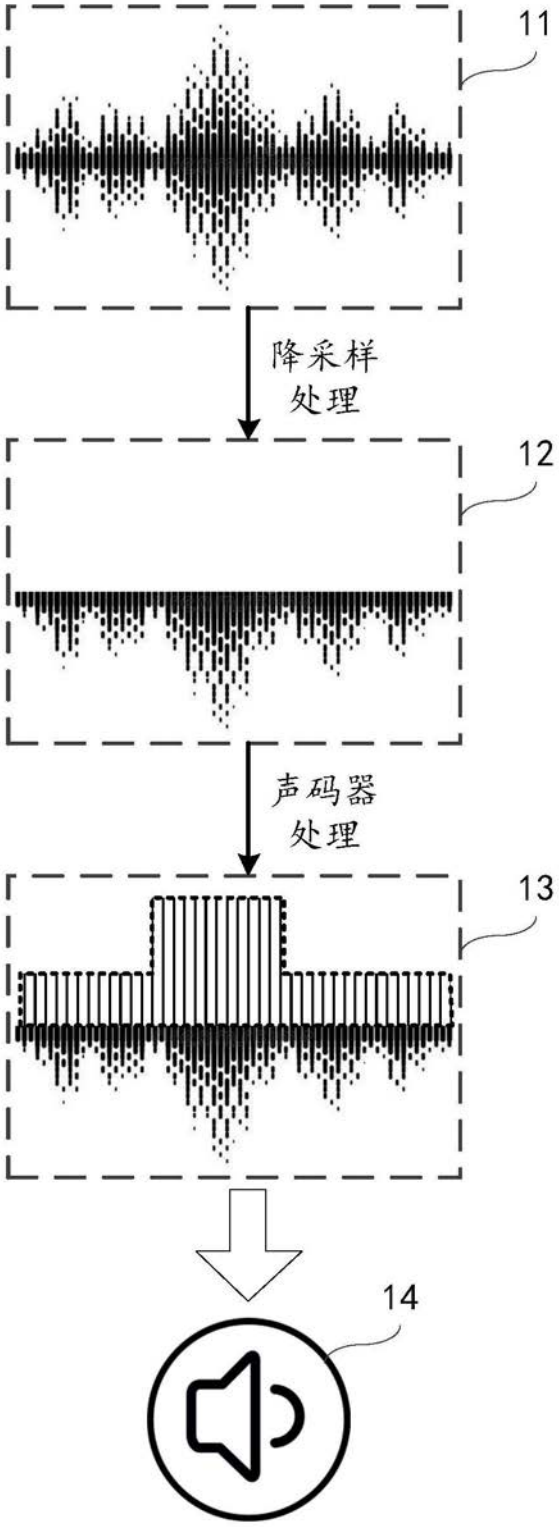

图1

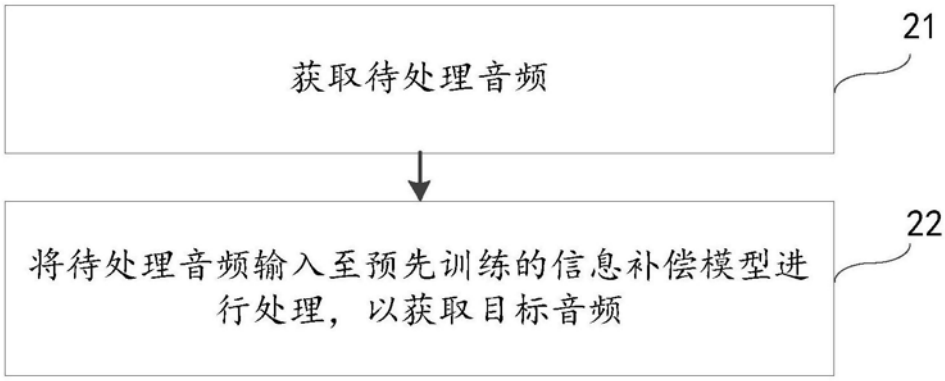
获取待处理音频
21
将待处理音频输入至预先训练的信息补偿模型进行处理，以获取目标音频
22

图2

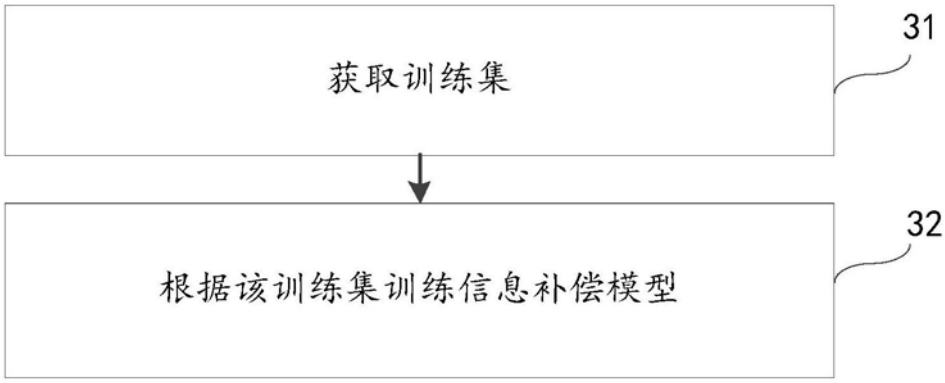
获取训练集
31
根据该训练集训练信息补偿模型
32

图3

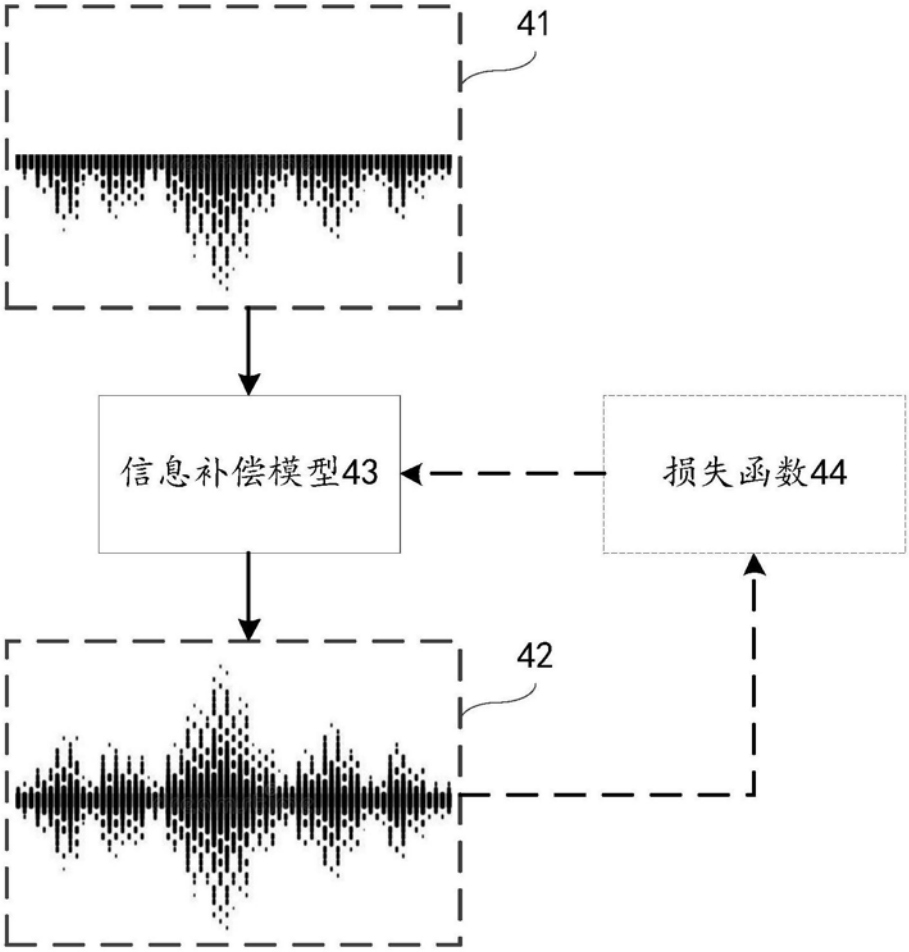

图4

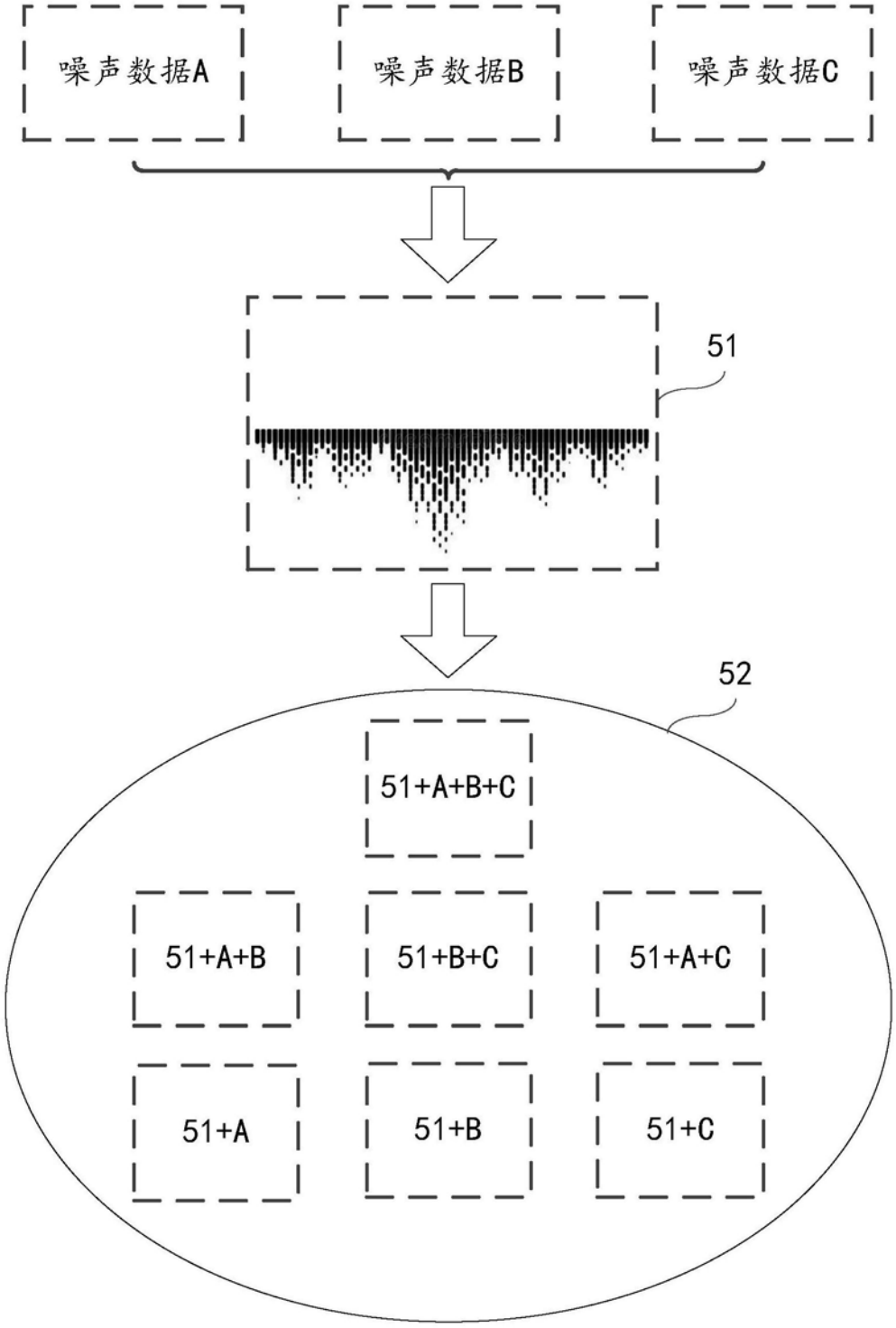

图5

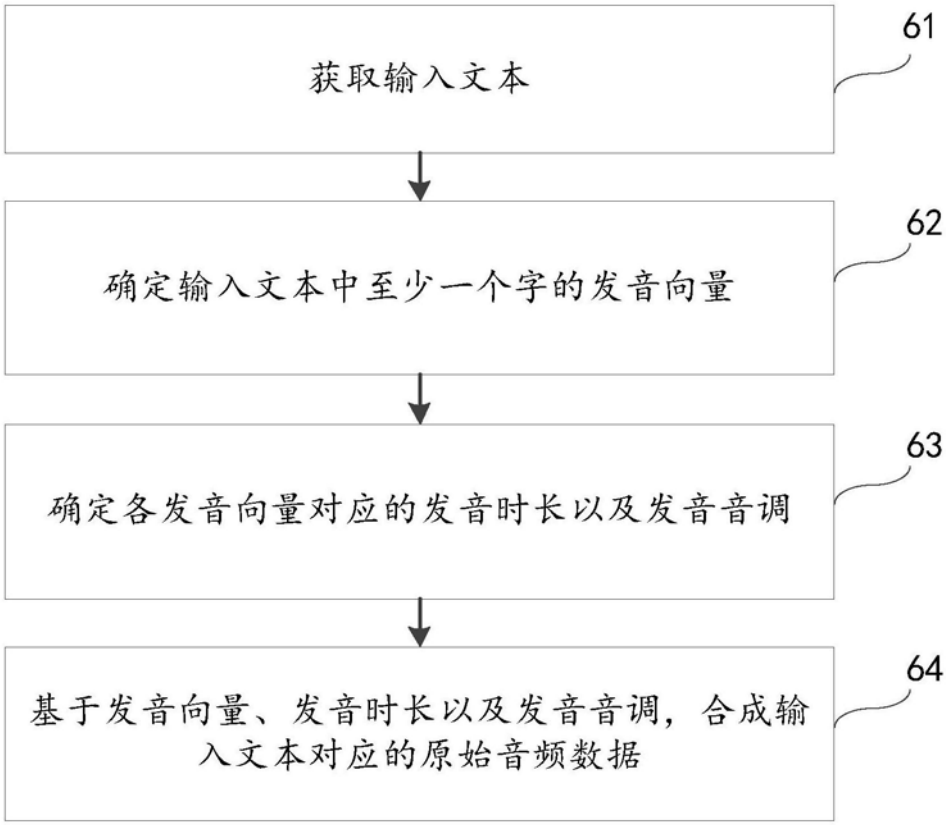
获取输入文本
61
确定输入文本中至少一个字的发音向量
62
确定各发音向量对应的发音时长以及发音音调
63
基于发音向量、发音时长以及发音音调，合成输入文本对应的原始音频数据
64

图6

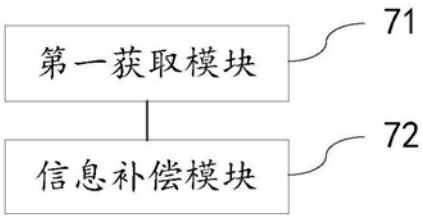
第一获取模块
71
信息补偿模块
72

图7

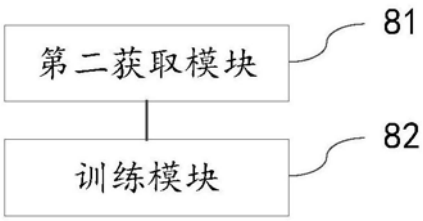
第二获取模块
81
训练模块
82

图8

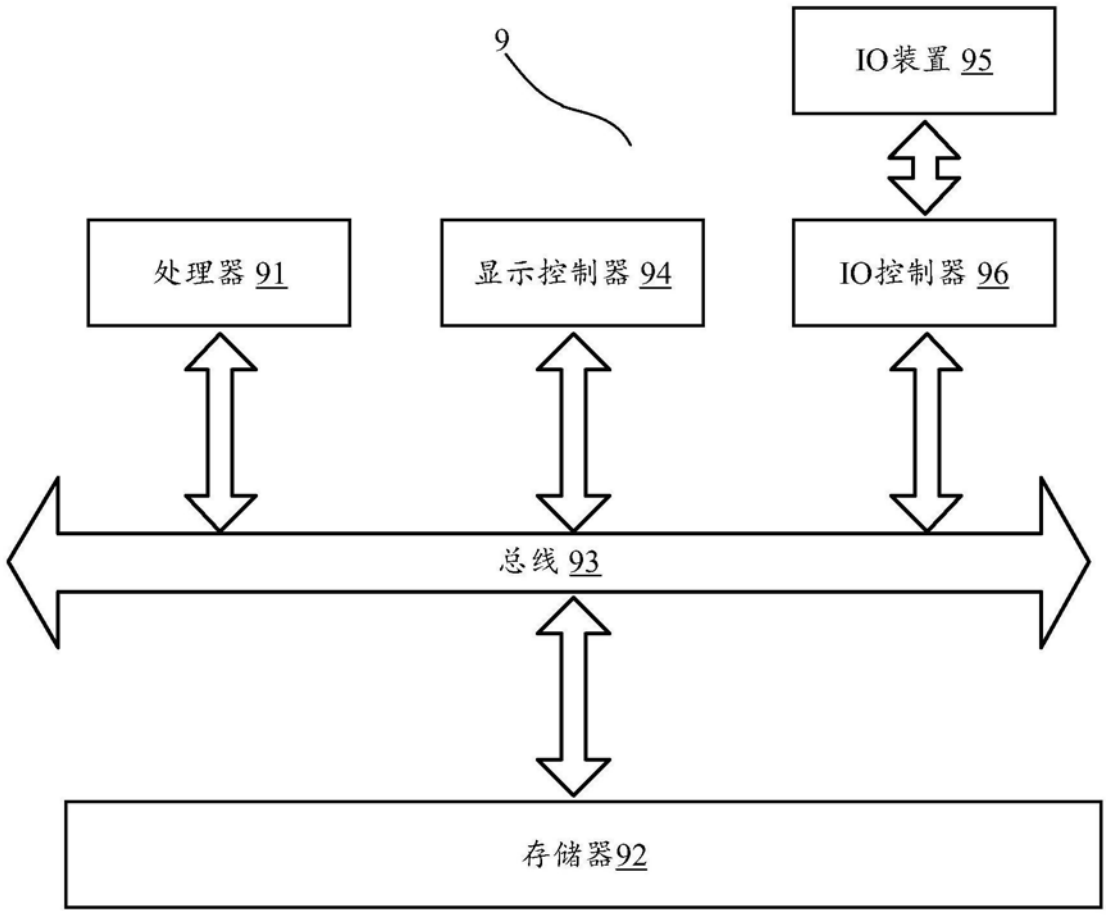

图9