(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 112712813 B
(45)授权公告日 2021.07.20

(21)申请号 202110327534.5
(22)申请日 2021.03.26
(65)同一申请的已公布的文献号
申请公布号 CN 112712813 A
(43)申请公布日 2021.04.27
(73)专利权人 北京达佳互联信息技术有限公司
地址 100085 北京市海淀区上地西路6号1幢1层101D1-7
(72)发明人 张颖
(74)专利代理机构 广州三环专利商标代理有限公司 44202
代理人 郝传鑫 贾允
(51)Int.Cl.
*G10L 21/013*(2013.01)

(56)对比文件
CN 110136690 A,2019.08.16
CN 112466316 A,2021.03.09
CN 111899719 A,2020.11.06
CN 109637551 A,2019.04.16
US 2020/0243101 A1,2020.07.30

审查员 王俊峰

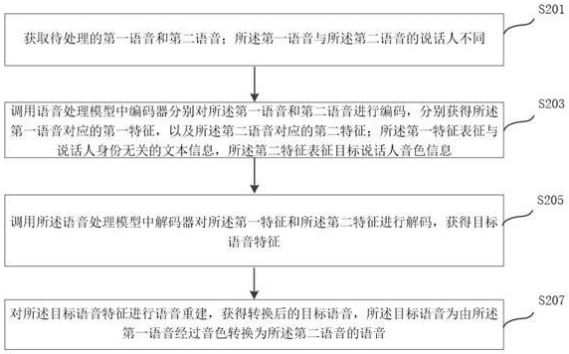
权利要求书5页 说明书16页 附图4页

(54)发明名称

# 语音处理方法、装置、设备及存储介质

(57)摘要

本公开关于一种语音处理方法、装置、设备及存储介质，该方法通过获取待处理的第一语音和第二语音；调用基于至少一句目标说话人语句进行优化训练得到的语音处理模型中编码器分别对获取的语音进行编码，分别获得表征与说话人身份无关的文本信息的第一特征和表征目标说话人音色信息的第二特征；基于第一特征和第二特征进行解码和语音重建，获得经过音色转换后的目标语音。如此，通过一个端到端的语音处理模型，使得语音处理模型无需大量的目标说话人语句，仅需基于少量话语就可完成目标说话人的音色建模能力，减少模型训练的计算资源的占用和耗时。




CN 112712813 B

1.一种语音处理方法,其特征在于,包括:

获取待处理的第一语音和第二语音;所述第一语音与所述第二语音的说话人不同;

调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码,分别获得所述第一语音对应的第一特征,以及所述第二语音对应的第二特征;所述第一特征表征与说话人身份无关的文本信息,所述第二特征表征目标说话人音色信息;所述语音处理模型是利用至少一句目标说话人语句对基础语音处理模型进行优化训练得到;其中,所述第二特征是根据说话人参考特征表示与先验说话人音色特征矩阵之间的相似度,以及所述说话人参考特征表示确定;所述说话人参考特征是基于所述编码器中参考信息编码子模块确定,且用于表征第二语音的说话人身份特性;所述先验说话人音色特征矩阵是基于所述编码器中先验音色信息子模块确定,且由多个先验说话人的身份特征向量所构成的矩阵;

调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码,获得目标语音特征;

对所述目标语音特征进行语音重建,获得转换后的目标语音,所述目标语音为由所述第一语音经过音色转换为所述第二语音的语音。

2.根据权利要求1所述的语音处理方法,其特征在于,所述编码器包括语言编码模块和音色编码模块;所述调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码,分别获得所述第一语音对应的第一特征,以及所述第二语音对应的第二特征步骤,包括:

调用所述语音处理模型中语言编码模块对所述第一语音进行编码,获得所述第一语音对应的第一特征;

调用所述语音处理模型中音色编码模块对所述第二语音进行编码,获得所述第二语音对应的第二特征。

3.根据权利要求2所述的语音处理方法,其特征在于,所述调用所述语音处理模型中语言编码模块对所述第一语音进行编码,获得所述第一语音对应的第一特征步骤,包括:

将所述第一语音输入至所述语音处理模型中语言编码模块;

利用所述语言编码模块对所述第一语音进行音素序列识别,得到多个音素序列;

利用所述语言编码模块计算每个音素序列对应语音类别的后验概率,获得每个音素序列对应的声学后验概率特征;

将多个音色序列对应的所述声学后验概率特征作为所述第一语音对应的第一特征。

4.根据权利要求2或3所述的语音处理方法,其特征在于,所述音色编码模块包括参考信息编码子模块、多头注意力子模块和先验音色信息子模块;所述调用所述语音处理模型中音色编码模块对所述第二语音进行编码,获得所述第二语音对应的第二特征步骤,包括:

提取所述第二语音的频谱特征;

将所述频谱特征输入至所述语音处理模型中参考信息编码子模块,利用所述参考信息编码子模块对所述频谱特征进行编码,获得说话人参考特征表示;

基于先验音色信息子模块,获取先验说话人音色特征矩阵;

利用所述语音处理模型中多头注意力子模块,计算所述说话人参考特征表示和所述先验说话人音色特征矩阵之间的相似度,获得目标说话人特征表示;

将所述目标说话人特征表示作为所述第二语音对应的第二特征。

5.根据权利要求4所述的语音处理方法,其特征在于,所述利用所述语音处理模型中多头注意力子模块,计算所述说话人参考特征表示和所述先验说话人音色特征矩阵之间的相似度,获得目标说话人特征表示步骤,包括:

分别对所述说话人参考特征表示和所述先验说话人音色特征矩阵进行维度规整,对应获得第一规整特征和第二规整特征;

分别将所述第一规整特征和所述第二规整特征进行分解,对应获得M个第一分解特征和M个第二分解特征;每个第一分解特征和每个第二分解特征分别对应一个注意力网络头的注意力子模块,其中,M为多头注意力子模块的头数;

针对每个注意力子模块,计算所述第一分解特征与所述第二分解特征的相似度,并基于计算得到的相似度和对应的第一分解特征,得到一个说话人向量表示;

拼接多个说话人向量表示,得到目标说话人特征表示。

6.根据权利要求1-3任一所述的语音处理方法,其特征在于,所述解码器包括全连接层、门控循环神经网络和输出层;所述调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码,获得目标语音特征步骤,包括:

调用所述解码器中全连接层对所述第一特征进行变换,得到全连接特征;

将所述全连接特征与所述第二特征进行拼接处理,得到拼接特征;

调用所述解码器中门控循环神经网络对所述拼接特征进行特征提取,得到时间相关特征;

将所述时间相关特征输入到所述解码器中输出层,得到目标语音特征。

7.根据权利要求1-3任一所述的语音处理方法,其特征在于,所述对所述目标语音特征进行语音重建,获得转换后的目标语音步骤,包括:

调用声码器对所述目标语音特征进行波形重建,获得转换后的目标语音。

8.根据权利要求1-3任一所述的语音处理方法,其特征在于,所述语音处理模型通过以下方式训练得到:

获取训练集,所述训练集包括至少一个语音样本对,每个语音样本对包括第一语音样本和第二语音样本,所述第一语音样本和第二语音样本为同一说话人的不同话语;

调用基础语音处理模型中编码器,分别对每个语音样本中第一语音样本和第二语音样本对进行编码,分别获得所述第一语音样本对对应的第一样本特征,以及所述第二语音样本对应的第二样本特征;所述第一样本特征表征与说话人身份无关的文本信息,所述第二样本特征表征目标说话人音色信息;

调用基础语音处理模型中解码器对每个语音样本对对应的第一样本特征和第二样本特征进行解码,获得每个语音样本对对应的目标语音样本特征;

对每个语音样本对中第一语音样本进行语音特征提取,获得实际目标语音特征;

计算每个语音样本对的目标语音样本特征与实际目标语音特征,确定基础语音处理模型的损失函数;

根据所述损失函数对所述基础语音处理模型进行训练,获得经训练的语音处理模型。

9.一种语音处理装置,其特征在于,包括:

获取模块,被配置为执行获取待处理的第一语音和第二语音;所述第一语音与所述第二语音的说话人不同;

编码模块，被配置为执行调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码，分别获得所述第一语音对应的第一特征，以及所述第二语音对应的第二特征；所述第一特征表征与说话人身份无关的文本信息，所述第二特征表征目标说话人音色信息；所述语音处理模型是利用至少一句目标说话人语句对基础语音处理模型进行优化训练得到；其中，所述第二特征是根据说话人参考特征表示与先验说话人音色特征矩阵之间的相似度，以及所述说话人参考特征表示确定；所述说话人参考特征是基于所述编码器中参考信息编码子模块确定，且用于表征第二语音的说话人身份特性；所述先验说话人音色特征矩阵是基于所述编码器中先验音色信息子模块确定，且由多个先验说话人的身份特征向量所构成的矩阵；

解码模块，被配置为执行调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码，获得目标语音特征；

重建模块，被配置为执行对所述目标语音特征进行语音重建，获得转换后的目标语音，所述目标语音为由所述第一语音经过音色转换为所述第二语音的语音。

10.根据权利要求9所述的语音处理装置，其特征在于，所述编码器包括语言编码模块和音色编码模块；所述编码模块包括：

第一编码子模块，被配置为执行调用所述语音处理模型中语言编码模块对所述第一语音进行编码，获得所述第一语音对应的第一特征；

第二编码子模块，被配置为执行调用所述语音处理模型中音色编码模块对所述第二语音进行编码，获得所述第二语音对应的第二特征。

11.根据权利要求10所述的语音处理装置，其特征在于，所述第一编码子模块包括：

输入单元，被配置为执行将所述第一语音输入至所述语音处理模型中语言编码模块；

识别单元，被配置为执行利用所述语言编码模块对所述第一语音进行音素序列识别，得到多个音素序列；

后验概率确定单元，被配置为执行利用所述语言编码模块计算每个音素序列对应语音类别的后验概率，获得每个音素序列对应的声学后验概率特征；

第一确定单元，被配置为执行将多个音色序列对应的所述声学后验概率特征作为所述第一语音对应的第一特征。

12.根据权利要求10或11所述的语音处理装置，其特征在于，所述音色编码模块包括参考信息编码子模块、多头注意力子模块和先验音色信息子模块；所述第二编码子模块包括：

特征提取单元，被配置为执行提取所述第二语音的频谱特征；

编码单元，被配置为执行将所述频谱特征输入至所述语音处理模型中参考信息编码子模块，利用所述参考信息编码子模块对所述频谱特征进行编码，获得说话人参考特征表示；

矩阵获取单元，被配置为执行基于先验音色信息子模块，获取先验说话人音色特征矩阵；

计算单元，被配置为执行利用所述语音处理模型中多头注意力子模块，计算所述说话人参考特征表示和所述先验说话人音色特征矩阵之间的相似度，获得目标说话人特征表示；

第二确定单元，被配置为执行将所述目标说话人特征表示作为所述第二语音对应的第二特征。

13.根据权利要求12所述的语音处理装置,其特征在于,所述计算单元包括:

规整子单元,被配置为执行分别对所述说话人参考特征表示和所述先验说话人音色特征矩阵进行维度规整,对应获得第一规整特征和第二规整特征;

分解子单元,被配置为执行分别将所述第一规整特征和所述第二规整特征进行分解,对应获得M个第一分解特征和M个第二分解特征;每个第一分解特征和每个第二分解特征分别对应一个注意力网络头的注意力子模块,其中,M为多头注意力子模块的头数;

计算子单元,被配置为执行针对每个注意力子模块,计算所述第一分解特征与所述第二分解特征的相似度,并基于计算得到的相似度和对应的第一分解特征,得到一个说话人向量表示;

拼接子单元,被配置为执行拼接多个说话人向量表示,得到目标说话人特征表示。

14.根据权利要求9-11任一所述的语音处理装置,其特征在于,所述解码器包括全连接层、门控循环神经网络和输出层;所述解码模块包括:

变换子模块,被配置为执行调用所述解码器中全连接层对所述第一特征进行变换,得到全连接特征;

拼接子模块,被配置为执行将所述全连接特征与所述第二特征进行拼接处理,得到拼接特征;

特征提取子模块,被配置为执行调用所述解码器中门控循环神经网络对所述拼接特征进行特征提取,得到时间相关特征;

预测子模块,被配置为执行将所述时间相关特征输入到所述解码器中输出层,得到目标语音特征。

15.根据权利要求9-11任一所述的语音处理装置,其特征在于,所述重建模块包括:

重建子模块,被配置为执行调用声码器对所述目标语音特征进行波形重建,获得转换后的目标语音。

16.根据权利要求9-11任一所述的语音处理装置,其特征在于,所述语音处理模型通过以下方式训练得到:

获取训练集,所述训练集包括至少一个语音样本对,每个语音样本对包括第一语音样本和第二语音样本,所述第一语音样本和第二语音样本为同一说话人的不同话语;

调用基础语音处理模型中编码器,分别对每个语音样本中第一语音样本和第二语音样本对进行编码,分别获得所述第一语音样本对对应的第一样本特征,以及所述第二语音样本对应的第二样本特征;所述第一样本特征表征与说话人身份无关的文本信息,所述第二样本特征表征目标说话人音色信息;

调用基础语音处理模型中解码器对每个语音样本对对应的第一样本特征和第二样本特征进行解码,获得每个语音样本对对应的目标语音样本特征;

对每个语音样本对中第一语音样本进行语音特征提取,获得实际目标语音特征;

计算每个语音样本对的目标语音样本特征与实际目标语音特征,确定基础语音处理模型的损失函数;

根据所述损失函数对所述基础语音处理模型进行训练,获得经训练的语音处理模型。

17.一种电子设备,其特征在于,包括:

处理器;

用于存储所述处理器可执行指令的存储器；

其中，所述处理器被配置为执行所述指令，以实现如权利要求1至8中任一项所述的语音处理方法。

18.一种计算机可读存储介质，当所述计算机可读存储介质中的指令由电子设备的处理器执行时，使得电子设备能够执行如权利要求1至8中任一项所述的语音处理方法。

# 语音处理方法、装置、设备及存储介质

## 技术领域

[0001] 本公开涉及计算机技术领域，尤其涉及一种语音处理方法、装置、设备及存储介质。

## 背景技术

[0002] 语音转换是指在保持语言内容不变的情况下，将语音的原始说话人音色转换为目标说话人的音色。语音转换在视频变声、视频配音、人机交互等领域均发挥着重要作用。

[0003] 相关技术中，现有的语音识别系统通常是采用大量的数据集训练而得。当目标说话人发生改变时，就需要获取大量的数据来重新训练一个语音转换模型，如此不仅耗费大量的计算机资源和时间，而且在有些特殊场景下，尤其是在新的目标说话人的语音数据比较少的情况下，不足以重新训练一个到新的目标说话人的语音转换模型。

## 发明内容

[0004] 本公开提供一种语音处理方法、装置、设备及存储介质，以至少解决相关技术中在目标说话人发生改变时，需要重新训练语音转换模型所带来的耗费大量的计算机资源和时间的问题，以及由于训练数据缺失而不足以重新训练语音转换模型等至少一种问题。本公开的技术方案如下：

[0005] 根据本公开实施例的第一方面，提供一种语音处理方法，包括：

[0006] 获取待处理的第一语音和第二语音；所述第一语音与所述第二语音的说话人不同；

[0007] 调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码，分别获得所述第一语音对应的第一特征，以及所述第二语音对应的第二特征；所述第一特征表征与说话人身份无关的文本信息，所述第二特征表征目标说话人音色信息；所述语音处理模型是利用至少一句目标说话人语句对基础语音处理模型进行优化训练得到；

[0008] 调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码，获得目标语音特征；

[0009] 对所述目标语音特征进行语音重建，获得转换后的目标语音，所述目标语音为由所述第一语音经过音色转换为所述第二语音的语音。

[0010] 作为一种可选的实施方式，所述编码器包括语言编码模块和音色编码模块；所述调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码，分别获得所述第一语音对应的第一特征，以及所述第二语音对应的第二特征步骤，包括：

[0011] 调用所述语音处理模型中语言编码模块对所述第一语音进行编码，获得所述第一语音对应的第一特征；

[0012] 调用所述语音处理模型中音色编码模块对所述第二语音进行编码，获得所述第二语音对应的第二特征。

[0013] 作为一种可选的实施方式，所述调用所述语音处理模型语言编码模块对所述第一

语音进行编码，获得所述第一语音对应的第一特征步骤，包括：

[0014] 将所述第一语音输入至所述语音处理模型中语言编码模块；

[0015] 利用所述语言编码模块对所述第一语音进行音素序列识别，得到多个音素序列；

[0016] 利用所述语言编码模块计算每个音素序列对应语音类别的后验概率，获得每个音素序列对应的声学后验概率特征；

[0017] 将多个音色序列对应的所述声学后验概率特征作为所述第一语音对应的第一特征。

[0018] 作为一种可选的实施方式，所述音色编码模块包括参考信息编码子模块、多头注意力子模块和先验音色信息子模块；所述调用所述语音处理模型中音色编码模块对所述第二语音进行编码，获得所述第二语音对应的第二特征步骤，包括：

[0019] 提取所述第二语音的频谱特征；

[0020] 将所述频谱特征输入至所述语音处理模型中参考信息编码子模块，利用所述参考信息编码子模块对所述频谱特征进行编码，获得说话人参考特征表示；

[0021] 基于先验音色信息子模块，获取先验说话人音色特征矩阵；

[0022] 利用所述语音处理模型中多头注意力子模块，计算所述说话人参考特征表示和所述先验说话人音色特征矩阵之间的相似度，获得目标说话人特征表示；

[0023] 将所述目标说话人特征表示作为所述第二语音对应的第二特征。

[0024] 作为一种可选的实施方式，所述利用所述语音处理模型中多头注意力子模块，计算所述说话人参考特征表示和所述先验说话人音色特征矩阵之间的相似度，获得目标说话人特征表示步骤，包括：

[0025] 分别对所述说话人参考特征表示和所述先验说话人音色特征矩阵进行维度规整，对应获得第一规整特征和第二规整特征；

[0026] 分别将所述第一规整特征和所述第一规整特征进行分解，对应获得M个第一分解特征和M个第二分解特征；每个第一分解特征和每个第二分解特征分别对应一个注意力网络头的注意力子模块，其中，M为多头注意力子模块的头数；

[0027] 针对每个注意力子模块，计算所述第一分解特征与所述第二分解特征的相似度，并基于计算得到的相似度和对应的第一分解特征，得到一个说话人向量表示；

[0028] 拼接多个说话人向量表示，得到目标说话人特征表示。

[0029] 作为一种可选的实施方式，所述解码器包括全连接层、门控循环神经网络和输出层；所述调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码，获得目标语音特征步骤，包括：

[0030] 调用所述解码器中全连接层对所述第一特征进行变换，得到全连接特征；

[0031] 将所述全连接特征与所述第二特征进行拼接处理，得到拼接特征；

[0032] 调用所述解码器中门控循环神经网络对所述拼接特征进行特征提取，得到时间相关特征；

[0033] 将所述时间相关特征输入到所述解码器中输出层，得到目标语音特征。

[0034] 作为一种可选的实施方式，所述对所述目标语音特征进行语音重建，获得转换后的目标语音步骤，包括：

[0035] 调用声码器对所述目标语音特征进行波形重建，获得转换后的目标语音。

[0036] 作为一种可选的实施方式，所述语音处理模型通过以下方式训练得到：

[0037] 获取训练集，所述训练集包括至少一个语音样本对，每个语音样本对包括第一语音样本和第二语音样本，所述第一语音样本和第二语音样本为同一说话人的不同话语；

[0038] 调用基础语音处理模型中编码器，分别对每个语音样本中第一语音样本和第二语音样本对进行编码，分别获得所述第一语音样本对对应的第一样本特征，以及所述第二语音样本对应的第二样本特征；所述第一样本特征表征与说话人身份无关的文本信息，所述第二样本特征表征目标说话人音色信息；

[0039] 调用基础语音处理模型中解码器对每个语音样本对对应的第一样本特征和第二样本特征进行解码，获得每个语音样本对对应的目标语音样本特征；

[0040] 对每个语音样本对中第一语音样本进行语音特征提取，获得实际目标语音特征；

[0041] 计算每个语音样本对的目标语音样本特征与实际目标语音特征，确定基础语音处理模型的损失函数；

[0042] 根据所述损失函数对所述基础语音处理模型进行训练，获得经训练的语音处理模型。

[0043] 根据本公开实施例的第二方面，提供一种语音处理装置，包括：

[0044] 获取模块，被配置为执行获取待处理的第一语音和第二语音；所述第一语音与所述第二语音的说话人不同；

[0045] 编码模块，被配置为执行调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码，分别获得所述第一语音对应的第一特征，以及所述第二语音对应的第二特征；所述第一特征表征与说话人身份无关的文本信息，所述第二特征表征目标说话人音色信息；所述语音处理模型是利用至少一句目标说话人语句对基础语音处理模型进行优化训练得到；

[0046] 解码模块，被配置为执行调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码，获得目标语音特征；

[0047] 重建模块，被配置为执行对所述目标语音特征进行语音重建，获得转换后的目标语音，所述目标语音为由所述第一语音经过音色转换为所述第二语音的语音。

[0048] 作为一种可选的实施方式，所述编码器包括语言编码模块和音色编码模块；所述编码模块包括：

[0049] 第一编码子模块，被配置为执行调用所述语音处理模型中语言编码模块对所述第一语音进行编码，获得所述第一语音对应的第一特征；

[0050] 第二编码子模块，被配置为执行调用所述语音处理模型中音色编码模块对所述第二语音进行编码，获得所述第二语音对应的第二特征。

[0051] 作为一种可选的实施方式，所述第一编码子模块包括：

[0052] 输入单元，被配置为执行将所述第一语音输入至所述语音处理模型中语言编码模块；

[0053] 识别单元，被配置为执行利用所述语言编码模块对所述第一语音进行音素序列识别，得到多个音素序列；

[0054] 后验概率确定单元，被配置为执行利用所述语言编码模块计算每个音素序列对应语音类别的后验概率，获得每个音素序列对应的声学后验概率特征；

[0055] 第一确定单元，被配置为执行将多个音色序列对应的所述声学后验概率特征作为所述第一语音对应的第一特征。

[0056] 作为一种可选的实施方式，所述音色编码模块包括参考信息编码子模块、多头注意力子模块和先验音色信息子模块。所述第二编码子模块包括：

[0057] 特征提取单元，被配置为执行提取所述第二语音的频谱特征；

[0058] 编码单元，被配置为执行将所述频谱特征输入至所述语音处理模型中参考信息编码子模块，利用所述参考信息编码子模块对所述频谱特征进行编码，获得说话人参考特征表示；

[0059] 矩阵获取单元，被配置为执行基于先验音色信息子模块，获取先验说话人音色特征矩阵；

[0060] 计算单元，被配置为执行利用所述语音处理模型中多头注意力子模块，计算所述说话人参考特征表示和所述先验说话人音色特征矩阵之间的相似度，获得目标说话人特征表示；

[0061] 第二确定单元，被配置为执行将所述目标说话人特征表示作为所述第二语音对应的第二特征。

[0062] 作为一种可选的实施方式，所述计算单元包括：

[0063] 规整子单元，被配置为执行分别对所述说话人参考特征表示和所述先验说话人音色特征矩阵和进行维度规整，对应获得第一规整特征和第二规整特征；

[0064] 分解子单元，被配置为执行分别将所述第一规整特征和所述第一规整特征进行分解，对应获得M个第一分解特征和M个第二分解特征；每个第一分解特征和每个第二分解特征分别对应一个注意力网络头的注意力子模块，其中，M为多头注意力子模块的头数；

[0065] 计算子单元，被配置为执行针对每个注意力子模块，计算所述第一分解特征与所述第二分解特征的相似度，并基于计算得到的相似度和对应的第一分解特征，得到一个说话人向量表示；

[0066] 拼接子单元，被配置为执行拼接多个说话人向量表示，得到目标说话人特征表示。

[0067] 作为一种可选的实施方式，所述解码器包括全连接层、门控循环神经网络和输出层；所述解码模块包括：

[0068] 变换子模块，被配置为执行调用所述解码器中全连接层对所述第一特征进行变换，得到全连接特征；

[0069] 拼接子模块，被配置为执行将所述全连接特征与所述第二特征进行拼接处理，得到拼接特征；

[0070] 特征提取子模块，被配置为执行调用所述解码器中门控循环神经网络对所述拼接特征进行特征提取，得到时间相关特征；

[0071] 预测子模块，被配置为执行将所述时间相关特征输入到所述解码器中输出层，得到目标语音特征。

[0072] 作为一种可选的实施方式，所述重建模块包括：

[0073] 重建子模块，被配置为执行调用声码器对所述目标语音特征进行波形重建，获得转换后的目标语音。

[0074] 作为一种可选的实施方式，所述语音处理模型通过以下方式训练得到：

[0075] 获取训练集，所述训练集包括至少一个语音样本对，每个语音样本对包括第一语音样本和第二语音样本，所述第一语音样本和第二语音样本为同一说话人的不同话语；

[0076] 调用基础语音处理模型中编码器，分别对每个语音样本中第一语音样本和第二语音样本对进行编码，分别获得所述第一语音样本对对应的第一样本特征，以及所述第二语音样本对应的第二样本特征；所述第一样本特征表征与说话人身份无关的文本信息，所述第二样本特征表征目标说话人音色信息；

[0077] 调用基础语音处理模型中解码器对每个语音样本对对应的第一样本特征和第二样本特征进行解码，获得每个语音样本对对应的目标语音样本特征；

[0078] 对每个语音样本对中第一语音样本进行语音特征提取，获得实际目标语音特征；

[0079] 计算每个语音样本对的目标语音样本特征与实际目标语音特征，确定基础语音处理模型的损失函数；

[0080] 根据所述损失函数对所述基础语音处理模型进行训练，获得经训练的语音处理模型。

[0081] 根据本公开实施例的第三方面，提供一种计算机可读存储介质，当所述计算机可读存储介质中的指令由电子设备的处理器执行时，使得电子设备能够执行如上述中任一实施方式所述的语音处理方法。

[0082] 根据本公开实施例的第四方面，提供一种电子设备，包括：

[0083] 处理器；

[0084] 用于存储所述处理器可执行指令的存储器；

[0085] 其中，所述处理器被配置为执行所述指令，以实现如上述任一实施方式所述的语音处理方法。

[0086] 根据本公开实施例的第五方面，提供一种计算机程序产品，该计算机程序产品包括计算机程序，该计算机程序被处理器执行时实现上述任一种实施方式中提供的语音处理方法。

[0087] 本公开的实施例提供的技术方案至少带来以下有益效果：

[0088] 本公开实施例通过获取待处理的第一语音和第二语音；所述第一语音与所述第二语音的说话人不同；调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码，分别获得所述第一语音对应的第一特征，以及所述第二语音对应的第二特征；所述语音处理模型是利用至少一句目标说话人语句对基础语音处理模型进行优化训练得到；所述第一特征表征与说话人身份无关的文本信息，所述第二特征表征目标说话人音色信息；调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码，获得目标语音特征；对所述目标语音特征进行语音重建，获得转换后的目标语音，所述目标语音为由所述第一语音经过音色转换为所述第二语音的语音。如此，通过一个端到端的语音处理模型，使得语音处理模型无需大量的目标说话人语句，仅需基于至少一句话语就可完成目标说话人的音色建模能力，也即基于少量话语就可完成目标说话人的音色建模能力，减少模型训练的计算资源的占用和耗时；同时，对于训练集中没有遇到的说话人，模型也能很好的预测其身份特征并完成语音转换，从而实现基于少样本甚至单样本学习到任意说话人的语音转换，提高了语音处理效率，降低了语音转换技术的应用门槛。

[0089] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不

能限制本公开。

## 附图说明

[0090] 此处的附图被并入说明书中并构成本公开的一部分，示出了符合本公开的实施例，并与说明书一起用于解释本公开的原理，并不构成对本公开的不当限定。

[0091] 图1是根据一示例性实施例示出的一种应用语音处理方法的系统的架构图。

[0092] 图2是根据一示例性实施例示出的一种语音处理方法的流程图。

[0093] 图3是根据一示例性实施例示出的一种语音处理方法的部分流程图。

[0094] 图4是根据一示例性实施例示出的一种获得目标说话人特征表示步骤的流程图。

[0095] 图5是根据一示例性实施例示出的语音处理模型的结构示意图。

[0096] 图6是根据一示例性实施例示出的一种语音处理装置框图。

[0097] 图7是根据一示例性实施例示出的一种电子设备的框图。

## 具体实施方式

[0098] 为了使本领域普通人员更好地理解本公开的技术方案，下面将结合附图，对本公开实施例中的技术方案进行清楚、完整地描述。

[0099] 需要说明的是，本公开的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本公开的实施例能够以除了在这里图示或描述的那些以外的顺序实施。以下示例性实施例中所描述的实施方式并不代表与本公开相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本公开的一些方面相一致的装置和方法的例子。

[0100] 首先，对本公开实施例中涉及的名词进行简单介绍：

[0101] 语音转换技术：是指保持语义内容不变的情况下，将源语音转换为目标语音的技术，其中，源语音为第一人声发出的语音，目标语音为第二人声发出的语音，也即将第一人声发出的源语音通过语音转换技术，转换为语义相同的第二人声发出的目标语音。

[0102] 音色：直译为声音的颜色、声音的色彩，是指声音的个性特征。音色的形成和差异是物体振动的不同分量组合变化关系在人耳的听觉上感受的效应。

[0103] 声学后验概率(Phonetic Posterior Grams，PPGs)：用于表达源语音的文本特征。

[0104] 图1是根据一示例性实施例示出的一种应用语音处理方法的系统的架构图，参见图1，该架构图可以包括终端10和服务器20。

[0105] 其中，终端10可以是但不限于为智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能可穿戴设备、数字助理、增强现实设备、虚拟现实设备等实体设备或者运行于实体设备中的应用程序、小程序中一种或多种。

[0106] 服务器20可以为终端提供语音处理等后台服务。服务器20可以响应于终端10发送的语音处理请求，获取第一语音(源语音)和第二语音(目标语音)，并对该第一语音和第二语音进行预测处理，得到目标语音特征；接着，基于该目标语音特征重建得到目标语音，也即得到由第一语音经过音色转换为第二语音的语音。

[0107] 仅作为示例，服务器20可以是但不限于为独立的服务器，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、中间服务、域名服务、安全服务、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等中一种或多种。该得终端10与服务器20之间可以通过有线或无线通信方式进行直接或间接地连接，本公开实施例在此不作限制。

[0108] 本公开实施例提供的语音处理方法，可以由语音处理装置来执行，该语音处理装置可以是硬件形式或软件形式集成在终端或服务器等电子设备中，也可以由终端或服务器单独执行实现，或者可以由终端和服务器协同执行实现。

[0109] 图2是根据一示例性实施例示出的一种语音处理方法的流程图，图5是根据一示例性实施例示出的语音处理模型的结构示意图。如图2和图5所示，语音处理方法可以应用于电子设备，以该电子设备为上述实施环境示意图中服务器为例进行说明，包括以下步骤。

[0110] 在步骤S201中，获取待处理的第一语音和第二语音；所述第一语音与所述第二语音的说话人不同。

[0111] 其中，第一语音为语音处理的源语音，也即需要进行语音转换处理的语音。该第一语音的获取方式可以包括：可以通过终端上的语音获取模块采集或录制得到、通过下载或手动输入等方式获取得到、或者通过从本地存储库或其他设备(例如云端或者服务器等)中获取得到等中至少一种，本公开对此不作具体限定。第一语音可以为说话语句、视频旁白、歌曲等类型，其数量可以为一个或多个。

[0112] 第二语音为目标语音，也即需要转换为目标音色对应的语音。所述第一语音与所述第二语音的说话人不同，也即第一语音对应的音色与第二语音对应的音色不同。该第二语音的获取方式也可以包括：可以通过终端上的语音获取模块采集或录制得到、通过下载或手动输入等方式获取得到、或者通过从本地存储库或其他设备(例如云端或者服务器等)中获取得到等中至少一种，本公开对此不作具体限定。第二语音可以为说话语句、视频旁白、歌曲等类型，该第二语音与第一语音的类型可以相同，也可不同；第二语音的数量优选为一个。

[0113] 可选地，服务器可以响应终端发送的语音处理指令，分别获取待处理的第一语音和第二语音，以调用语音处理模型对获取得到的第一语音和第二语音进行相应的语音处理。

[0114] 在步骤S203中，调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码，分别获得所述第一语音对应的第一特征，以及所述第二语音对应的第二特征；所述第一特征表征与说话人身份无关的文本信息，所述第二特征表征目标说话人音色信息。

[0115] 其中，所述语音处理模型是利用至少一句目标说话人语句对基础语音处理模型进行优化训练得到。这里的基础语音处理模型可以是利用多说话人基础数据库训练得到的具有语音处理功能的模型。仅作为示例，至少一句目标说话人语句是指少量目标说话人语句，其可以为几十句目标说话人语句、几句目标说话人语句，甚至一句目标说话人语句。由于基础语音处理模型是利用多说话人基础数据库训练得到，而该多说话人基础数据库中通常不包含目标说话人，故而无法准确提取目标说话人的音色信息相关特征。通过利用少量目标说话人语句对基础语音处理模型进行优化训练得到，使得训练得到的语音处理模型具有提取目标守护人的音色信息相关特征的能力。

[0116] 在一可选实施例中，所述编码器包括语言编码模块和音色编码模块。该语言编码

模块用于对语音中的语言内容信息进行编码，音色编码模块用于对语音中的音色内容信息进行编码。此时，上述步骤S203中，所述调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码，分别获得所述第一语音对应的第一特征，以及所述第二语音对应的第二特征步骤，包括：

[0117] 在步骤S2031中，调用所述语音处理模型中语言编码模块对所述第一语音进行编码，获得所述第一语音对应的第一特征。

[0118] 可选地，所述调用所述语音处理模型中语言编码模块对所述第一语音进行编码，获得所述第一语音对应的第一特征步骤，包括：将所述第一语音输入至所述语音处理模型中语言编码模块；利用所述语言编码模块对所述第一语音进行音素序列识别，得到多个音素序列；利用所述语言编码模块计算每个音素序列对应语音类别的后验概率，获得每个音素序列对应的声学后验概率特征；将多个音色序列对应的所述声学后验概率特征作为所述第一语音对应的第一特征。

[0119] 其中，声学后验概率(PPGs)特征是一个时间-类别的矩阵表示，其可以反映对于一个话语的每个特定时间帧，在每个语音类别的后验概率。该语音类别可以指词、音素或音素状态。

[0120] 上述实施例，通过将第一语音输入至语音处理模型中语言编码模块，先对第一语音进行音素序列识别，得到多个音素序列；再计算每个音素序列对应语音类别的后验概率，获得每个音素序列对应的声学后验概率特征；接着将多个音色序列对应的声学后验概率特征作为第一语音对应的第一特征。由于声学后验概率特征是一个与说话人无关的文本特征，能够更好的结合表征目标说话人音色信息的第二特征进行语音处理，提高语音处理质量。

[0121] 在步骤S2033中，调用所述语音处理模型中音色编码模块对所述第二语音进行编码，获得所述第二语音对应的第二特征。

[0122] 该音色编码模块通过将目标说话人的说话人特征表示投影至预先构建的说话人表示空间中，利用说话人表示空间对应的多个说话人级别表示矩阵来表征目标说话人特征表示，也即得到该第二语音对应的第二特征。

[0123] 在一可选实施例中，所述音色编码模块包括参考信息编码子模块、多头注意力子模块和先验音色信息子模块。

[0124] 其中，参考信息编码子模块用于将变长的语音信号压缩成定长的参考特征(reference embedding)表示。多头注意力子模块可以为多头注意力网络(multi-head attention net)。先验音色信息子模块基于现有声纹识别的多个x-vector(一种身份特征向量)特征所构成的特征矩阵。

[0125] 上述实施例，通过调用语音处理模型中编码器的语言编码模块和音色编码模块，分别对第一语音和第二语音进行编码，获得对应的第一特征和第二特征。由于第一特征表征与说话人身份无关的文本信息，第二特征表征目标说话人音色信息，通过两个编码模块分别对第一语音和第二语音进行单独编码控制，能够更好地提取不同的特征信息，提高了特征编码质量和效率，减少了语音处理耗时，提高了语音处理效果。

[0126] 此时，参见图3所示，所述调用所述语音处理模型中音色编码模块对所述第二语音进行编码，获得所述第二语音对应的第二特征步骤，包括：

[0127] 在步骤S301中，提取所述第二语音的频谱特征。

[0128] 可选地，服务器获取第二语音之后，可以将第二语音进行时频域转换，提取第二语音的频谱特征。该频谱特征可以包括对数梅尔频谱特征，其特征维度可以为$L_R \times d_R$。

[0129] 在步骤S303中，将所述频谱特征输入至所述语音处理模型中参考信息编码子模块，利用所述参考信息编码子模块对所述频谱特征进行编码，获得说话人参考特征表示。

[0130] 可选地，服务器获取第二语音的频谱特征之后，可以将该第二语音的频谱特征作为输入特征，输入至音色编码模块中参考信息编码子模块，利用该参考信息编码子模块对频谱特征进行编码，输出得到说话人参考特征表示。该说话人参考特征表示用于反映第二语音的说话人身份特性。该说话人参考特征表示可以是特征向量的表示形式，其特征维度可以为$d_p$。

[0131] 可选地，参考信息编码子模块可以包括卷积块、RNN(Recurrent Neural Network，循环神经网络)模块和全连接层。该卷积块可以包括6个二维卷积层，每个卷积层的卷积核尺寸可以为3×3，卷积步长可以为2×2。每层卷积层连批量归一化层和激活层，该激活层可以包括ReLU激活函数，最后一个卷积层的输出通道维度为128，这样输入的频谱特征经过该卷积块进行下采样到(LR/64)×128(dR/64)维度。将卷积块的输出经过RNN模块压缩为单个固定长度的向量。接着，将该RNN模块输出的128维向量输入至全连接层，得到指定维度$d_p$的说话人参考特征表示。仅作为示例，该RNN模块可以包括具有128个计算单元的门控循环单元(GRU)，该全连接层的激活函数可以为tanh。

[0132] 可选地，该说话人参考特征表示可以是句子级的特征表示。相比于帧级别的说话人特征，句子级的特征对语句内容的时间变化更不敏感，更适合表示全局稳定的说话人音色特征。

[0133] 在步骤S305中，基于先验音色信息子模块，获取先验说话人音色特征矩阵。

[0134] 可选地，先验音色信息子模块是基于现有声纹识别的多个x-vector特征所构成的特征矩阵$S$。特征矩阵S是从训练集筛选说话人的x-vector拼接而成的，在筛选说话人时考虑性别均衡(也即男女各一半)。在先验音色信息子模块中所有说话人的x-vector分布在一个高维的说话人表示空间中，每个x-vector可以看成该说话人表示空间的一个空间坐标点。这里，将先验音色信息子模块可以表示为$S=(S_1,...,S_N)$，其中N是基础说话人的总数，每个$S_i$表示一个说话人的x-vector，每个$S_i$的维度可以为200维。由于筛选的x-vector是经过现有的训练好的网络模型提取得到的特征，且提取得到的x-vector特征是与说话人身份密切相关的，故而对于一个新的说话人特征表示可以根据所有说话人的x-vector进行加权量化得到。

[0135] 在步骤S307中，利用所述语音处理模型中多头注意力子模块，计算所述说话人参考特征表示和所述先验说话人音色特征矩阵之间的相似度，获得目标说话人特征表示。

[0136] 在一可选实施例中，参见图4所示，所述利用所述语音处理模型中多头注意力子模块，计算所述先验说话人音色特征矩阵和所述说话人参考特征表示之间的相似度，获得目标说话人特征表示步骤，包括：

[0137] 在步骤S3071中，分别对所述说话人参考特征表示和所述先验说话人音色特征矩阵进行维度规整，对应获得第一规整特征和第二规整特征。

[0138] 可选地，第一规整特征$Z_q$和第二规整特征$Z_s$可以表示为：

[0139] $$Z_q = \omega_q \times z_{d_p}$$

[0140] $$Z_s = \omega_{kv} \times S$$

[0141] 其中，$\omega_q$和$\omega_{kv}$分别表示说话人参考特征表示以及先验说话人音色特征矩阵的权重系数，$Z_q$和$Z_s$分别是规整到维度为$d_q$和$d_{kv}$的向量。

[0142] 在步骤S3073中，分别将所述第一规整特征和所述第一规整特征进行分解，对应获得M个第一分解特征和M个第二分解特征；每个第一分解特征和每个第二分解特征分别对应一个注意力网络头的注意力子模块，其中，M为多头注意力子模块的头数，M可取正整数。

[0143] 可选地，$Z_q$和$Z_s$分被分解成$(Z_q^1,...,Z_q^M)$和$(Z_s^1,...,Z_s^M)$，其中M是注意力机制网络的头数，这些被分解的M个第一分解特征和M个第二分解特征被分配到对应的注意力子模块中用于后续计算。具体地，第i个注意力网络头的第一分解特征$Z_q^i$和第二分解特征$Z_s^i$分配到的第i个注意力网络头对应的注意力子模块中用于后续计算。

[0144] 在步骤S3075中，针对每个注意力子模块，计算所述第一分解特征与所述第二分解特征的相似度，并基于计算得到的相似度和对应的第一分解特征，得到一个说话人向量表示。

[0145] 可选地，一个说话人向量表示的计算公式可以表示为：

[0146] $$e_i = soft\max(\frac{Z_q^i \times Z_s^i}{A}) \times Z_q^i$$

[0147] 其中，$A = \sqrt{d_q / M}$，$e_i$表示第i个注意力网络头的说话人向量。

[0148] 在步骤S3077中，拼接多个说话人向量表示，得到目标说话人特征表示。

[0149] 可选地，目标说话人特征表示$speaker_embedding = [e_1, e_2, ..., e_M]$。

[0150] 上述实施例，分别对说话人参考特征表示和先验说话人音色特征矩阵和进行维度规整，对应获得第一规整特征和第二规整特征；之后，分别将第一规整特征和第一规整特征进行分解，对应获得M个第一分解特征和M个第二分解特征；每个第一分解特征和每个第二分解特征分别对应一个注意力网络头的注意力子模块，其中，M为多头注意力子模块的头数；针对每个注意力子模块，计算第一分解特征与第二分解特征的相似度，并基于计算得到的相似度和对应的第一分解特征，得到一个说话人向量表示；拼接多个说话人向量表示，得到目标说话人特征表示。从而可以提高第二特征的提取准确率，利于提高后续语音处理效果。

[0151] 在步骤S309中，将所述目标说话人特征表示作为所述第二语音对应的第二特征。

[0152] 上述实施例，通过将音色编码模块分为参考信息编码子模块、多头注意力子模块和先验音色信息子模块，并分别利用参考信息编码子模块进行频谱特征编码，获得说话人参考特征表示，利用先验音色信息子模块获取先验说话人音色特征矩阵，以及利用多头注意力子模块计算说话人参考特征表示与先验说话人音色特征矩阵的相似度，获得目标说话

人特征表示。由于先验说话人音色特征矩阵中包括多个已知的说话人音色特征信息，通过先验说话人音色特征矩阵与说话人参考特征表示的相似度，来获取作为第二语音对应的第二特征的目标说话人特征表示，可以有效地继承已经训练好的网络模型的特征提取能力，减少特征提取计算量和耗时，同时也提高了特征提取质量，利于提高后续语音处理效果。

[0153] 在步骤S205中，调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码，获得目标语音特征。

[0154] 在一可选实施例中，所述解码器可以包括全连接层、门控循环神经网络和输出层。此时，所述调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码，获得目标语音特征步骤，包括：

[0155] 在步骤S2051中，调用所述解码器中全连接层对所述第一特征进行变换，得到全连接特征；

[0156] 在步骤S2053中，将所述全连接特征与所述第二特征进行拼接处理，得到拼接特征；

[0157] 在步骤S2055中，调用所述解码器中门控循环神经网络对所述拼接特征进行特征提取，得到时间相关特征；

[0158] 在步骤S2057中，将所述时间相关特征输入到所述解码器中输出层，得到目标语音特征。

[0159] 可选地，将所述第一特征和所述第二特征作为所述语音处理模型中解码器的输入，以预测目标语音特征。该目标语音特征用于反映目标说话人的声学特征。

[0160] 在实际应用中，可以先将所述第一特征输入至解码器中的全连接层，然后，将全连接层的输出与第二特征在最后一维进行特征拼接，并一起送到门控循环神经网络，以建模与时间相关的特性。接着，再输入至输出层进行预测，得到目标语音特征。可选地，全连接层的单元数量为256个，控循环神经网络的数量为2个且包含256个单元，输出层包括一个包含20个单元的全连接层，预测得到的目标语音特征为20维与声码器相关的声学特征。

[0161] 上述实施例，通过调用解码器中全连接层对第一特征进行变换，得到全连接特征；接着将全连接特征与所述第二特征进行拼接处理，得到拼接特征；调用解码器中门控循环神经网络对拼接特征进行特征提取，得到时间相关特征；将时间相关特征输入到解码器中输出层，得到目标语音特征，使得输出的目标语音特征符合第一语音的文本内容，又符合第二语音的说话人音色信息，提高了语音处理效果。

[0162] 在步骤S207中，对所述目标语音特征进行语音重建，获得转换后的目标语音，所述目标语音为由所述第一语音经过音色转换为所述第二语音的语音。

[0163] 可选地，服务器获取到目标语音特征之后，可以调用声码器对所述目标语音特征进行波形重建，获得转换后的目标语音，也即转换为人耳能够识别的目标语音。该目标语音为由所述第一语音经过音色转换为所述第二语音的语音。其中，声码器包括但不限于为LPCNet、WORLD、STRAIGHT、WaveNet等中至少一种。

[0164] 在一可选实施方式中，所述语音处理模型通过以下方式训练得到：

[0165] 获取训练集，所述训练集包括至少一个语音样本对，每个语音样本对包括第一语音样本和第二语音样本，所述第一语音样本和第二语音样本为同一说话人的不同话语；

[0166] 调用基础语音处理模型中编码器，分别对每个语音样本中第一语音样本和第二语

音样本对进行编码，分别获得所述第一语音样本对对应的第一样本特征，以及所述第二语音样本对应的第二样本特征；所述第一样本特征表征与说话人身份无关的文本信息，所述第二样本特征表征目标说话人音色信息；

[0167] 调用基础语音处理模型中解码器对每个语音样本对对应的第一样本特征和第二样本特征进行解码，获得每个语音样本对对应的目标语音样本特征；

[0168] 对每个语音样本对中第一语音样本进行语音特征提取，获得实际目标语音特征；

[0169] 计算每个语音样本对的目标语音样本特征与实际目标语音特征，确定基础语音处理模型的损失函数；

[0170] 根据所述损失函数对所述基础语音处理模型进行训练，获得经训练的语音处理模型。

[0171] 其中，基础语音处理模型的损失函数可以为交叉熵损失函数。根据损失函数对所述基础语音处理模型进行反向传播，在解码器重建损失函数的指导下和其他部分的参数一同更新，不断优化网络参数，直至满足训练结束条件，获得经训练的语音处理模型。该训练结束条件可以包括但不限于为损失函数最小化、达到预设训练次数等。

[0172] 上述实施例，第一样本特征是一个说话人无关的文本特征，而语音处理模型的输出为目标语音样本特征，也即该文本特征对应的声学特征。由于第一样本特征是说话人无关的特征，而语音处理模型是在多说话人数据集上训练的，也即涵盖多个说话人身份，从而在损失函数的指导下，用于提取第二特征的部分编码器会朝着分离语句中说话人身份特征表示的方向更新，这样即使对于训练集中没有遇到的说话人，模型也能很好的预测其身份特征，并完成语音转换，同时也大大简化了模型训练过程。

[0173] 本公开实施例通过获取待处理的第一语音和第二语音；所述第一语音与所述第二语音的说话人不同；调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码，分别获得所述第一语音对应的第一特征，以及所述第二语音对应的第二特征；所述语音处理模型是利用至少一句目标说话人语句对基础语音处理模型进行优化训练得到；所述第一特征表征与说话人身份无关的文本信息，所述第二特征表征目标说话人音色信息；调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码，获得目标语音特征；对所述目标语音特征进行语音重建，获得转换后的目标语音，所述目标语音为由所述第一语音经过音色转换为所述第二语音的语音。如此，通过一个端到端的语音处理模型，使得语音处理模型无需大量的目标说话人语句，仅需基于至少一句话语就可完成目标说话人的音色建模能力，也即基于少量话语就可完成目标说话人的音色建模能力，减少模型训练的计算资源的占用和耗时；同时，对于训练集中没有遇到的说话人，模型也能很好的预测其身份特征并完成语音转换，从而实现基于少样本甚至单样本学习到任意说话人的语音转换，提高了语音处理效率，降低了语音转换技术的应用门槛。

[0174] 图6是根据一示例性实施例示出的一种语音处理装置框图。参照图6，该装置应用于电子设备，包括：

[0175] 获取模块610，被配置为执行获取待处理的第一语音和第二语音；所述第一语音与所述第二语音的说话人不同；

[0176] 编码模块620，被配置为执行调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码，分别获得所述第一语音对应的第一特征，以及所述第二语音对应的第

二特征;所述第一特征表征与说话人身份无关的文本信息,所述第二特征表征目标说话人音色信息;所述语音处理模型是基利用至少一句目标说话人语句对基础语音处理模型进行优化训练得到;

[0177] 解码模块630,被配置为执行调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码,获得目标语音特征;

[0178] 重建模块640,被配置为执行对所述目标语音特征进行语音重建,获得转换后的目标语音,所述目标语音为由所述第一语音经过音色转换为所述第二语音的语音。

[0179] 作为一种可选的实施方式,所述编码器包括语言编码模块和音色编码模块;所述编码模块包括:

[0180] 第一编码子模块,被配置为执行调用所述语音处理模型中语言编码模块对所述第一语音进行编码,获得所述第一语音对应的第一特征;

[0181] 第二编码子模块,被配置为执行调用所述语音处理模型中音色编码模块对所述第二语音进行编码,获得所述第二语音对应的第二特征。

[0182] 作为一种可选的实施方式,所述第一编码子模块包括:

[0183] 输入单元,被配置为执行将所述第一语音输入至所述语音处理模型中语言编码模块;

[0184] 识别单元,被配置为执行利用所述语言编码模块对所述第一语音进行音素序列识别,得到多个音素序列;

[0185] 后验概率确定单元,被配置为执行利用所述语言编码模块计算每个音素序列对应语音类别的后验概率,获得每个音素序列对应的声学后验概率特征;

[0186] 第一确定单元,被配置为执行将多个音色序列对应的所述声学后验概率特征作为所述第一语音对应的第一特征。

[0187] 作为一种可选的实施方式,所述音色编码模块包括参考信息编码子模块、多头注意力子模块和先验音色信息子模块。所述第二编码子模块包括:

[0188] 特征提取单元,被配置为执行提取所述第二语音的频谱特征;

[0189] 编码单元,被配置为执行将所述频谱特征输入至所述语音处理模型中参考信息编码子模块,利用所述参考信息编码子模块对所述频谱特征进行编码,获得说话人参考特征表示;

[0190] 矩阵获取单元,被配置为执行基于先验音色信息子模块,获取先验说话人音色特征矩阵;

[0191] 计算单元,被配置为执行利用所述语音处理模型中多头注意力子模块,计算所述说话人参考特征表示和所述先验说话人音色特征矩阵之间的相似度,获得目标说话人特征表示;

[0192] 第二确定单元,被配置为执行将所述目标说话人特征表示作为所述第二语音对应的第二特征。

[0193] 作为一种可选的实施方式,所述计算单元包括:

[0194] 规整子单元,被配置为执行分别对所述说话人参考特征表示和所述先验说话人音色特征矩阵进行维度规整,对应获得第一规整特征和第二规整特征;

[0195] 分解子单元,被配置为执行分别将所述第一规整特征和所述第一规整特征进行分

解，对应获得M个第一分解特征和M个第二分解特征；每个第一分解特征和每个第二分解特征分别对应一个注意力网络头的注意力子模块，其中，M为多头注意力子模块的头数；

[0196] 计算子单元，被配置为执行针对每个注意力子模块，计算所述第一分解特征与所述第二分解特征的相似度，并基于计算得到的相似度和对应的第一分解特征，得到一个说话人向量表示；

[0197] 拼接子单元，被配置为执行拼接多个说话人向量表示，得到目标说话人特征表示。

[0198] 作为一种可选的实施方式，所述解码器包括全连接层、门控循环神经网络和输出层；所述解码模块包括：

[0199] 变换子模块，被配置为执行调用所述解码器中全连接层对所述第一特征进行变换，得到全连接特征；

[0200] 拼接子模块，被配置为执行将所述全连接特征与所述第二特征进行拼接处理，得到拼接特征；

[0201] 特征提取子模块，被配置为执行调用所述解码器中门控循环神经网络对所述拼接特征进行特征提取，得到时间相关特征；

[0202] 预测子模块，被配置为执行将所述时间相关特征输入到所述解码器中输出层，得到目标语音特征。

[0203] 作为一种可选的实施方式，所述重建模块包括：

[0204] 重建子模块，被配置为执行调用声码器对所述目标语音特征进行波形重建，获得转换后的目标语音。

[0205] 作为一种可选的实施方式，所述语音处理模型通过以下方式训练得到：

[0206] 获取训练集，所述训练集包括至少一个语音样本对，每个语音样本对包括第一语音样本和第二语音样本，所述第一语音样本和第二语音样本为同一说话人的不同话语；

[0207] 调用基础语音处理模型中编码器，分别对每个语音样本中第一语音样本和第二语音样本对进行编码，分别获得所述第一语音样本对对应的第一样本特征，以及所述第二语音样本对应的第二样本特征；所述第一样本特征表征与说话人身份无关的文本信息，所述第二样本特征表征目标说话人音色信息；

[0208] 调用基础语音处理模型中解码器对每个语音样本对对应的第一样本特征和第二样本特征进行解码，获得每个语音样本对对应的目标语音样本特征；

[0209] 对每个语音样本对中第一语音样本进行语音特征提取，获得实际目标语音特征；

[0210] 计算每个语音样本对的目标语音样本特征与实际目标语音特征，确定基础语音处理模型的损失函数；

[0211] 根据所述损失函数对所述基础语音处理模型进行训练，获得经训练的语音处理模型。

[0212] 关于上述实施例中的装置，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

[0213] 在示例性实施例中，还提供了一种电子设备，电子设备包括处理器；用于存储处理器可执行指令的存储器；其中，处理器被配置为执行存储器上所存放的指令时，实现上述实施例中任一语音处理方法的步骤。

[0214] 该电子设备可以终端、服务器或者类似的运算装置，以该电子设备是服务器为例，

图7是根据一示例性实施例示出的一种用于语音处理的电子设备的框图，具体来讲：

[0215] 该电子设备1000可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上中央处理器(Central Processing Units,CPU)1010(处理器1010可以包括但不限于微处理器MCU或可编程逻辑器件FPGA等的处理装置)、用于存储数据的存储器1030，一个或一个以上存储应用程序1023或数据1022的存储介质1020(例如一个或一个以上海量存储设备)。其中，存储器1030和存储介质1020可以是短暂存储或持久存储。存储在存储介质1020的程序可以包括一个或一个以上模块，每个模块可以包括对电子设备中的一系列指令操作。更进一步地，中央处理器1010可以设置为与存储介质1020通信，在电子设备1000上执行存储介质1020中的一系列指令操作。

[0216] 电子设备1000还可以包括一个或一个以上电源1060，一个或一个以上有线或无线网络接口1050，一个或一个以上输入输出接口1040，和/或，一个或一个以上操作系统1021，例如Windows Server,Mac OS X,Unix, Linux,FreeBSD等等。

[0217] 输入输出接口1040可以用于经由一个网络接收或者发送数据。上述的网络具体实例可包括电子设备1000的通信供应商提供的无线网络。在一个实例中，输入输出接口1040包括一个网络适配器(Network Interface Controller,NIC)，其可通过基站与其他网络设备相连从而可与互联网进行通讯。在一个示例性实施例中，输入输出接口1040可以为射频(Radio Frequency,RF)模块，其用于通过无线方式与互联网进行通讯。

[0218] 本领域普通技术人员可以理解，图7所示的结构仅为示意，其并不对上述电子设备的结构造成限定。例如，电子设备1000还可包括比图7中所示更多或者更少的组件，或者具有与图7所示不同的配置。

[0219] 在示例性实施例中，还提供了一种计算机存储介质，当计算机存储介质中的指令由电子设备的处理器执行时，使得电子设备能够执行上述任一种实施方式中提供的方法的步骤。

[0220] 在示例性实施例中，还提供了一种计算机程序产品，该计算机程序产品包括计算机程序/指令，该所述计算机程序/指令被处理器执行时实现上述任一种实施方式中提供的方法。可选地，该计算机程序存储在计算机可读存储介质中。电子设备的处理器从计算机可读存储介质读取该计算机程序，处理器执行该计算机程序，使得该电子设备执行上述任一种实施方式中提供的方法。

[0221] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，该计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink) DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0222] 本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本公开的其

它实施方案。本申请旨在涵盖本公开的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本公开的真正范围和精神由下面的权利要求指出。

[0223] 应当理解的是，本公开并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本公开的范围仅由所附的权利要求来限制。

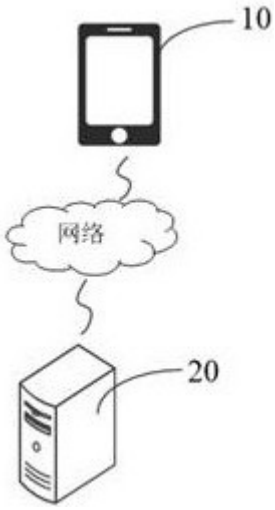

10
网络
20

图1

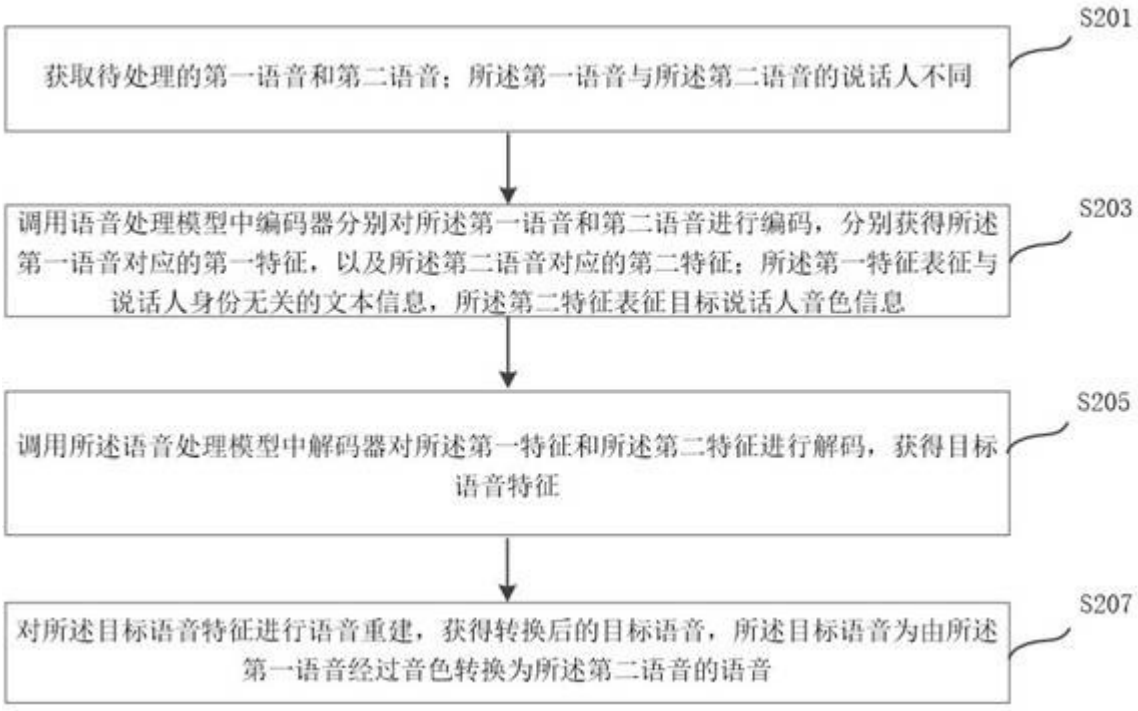

S201
获取待处理的第一语音和第二语音；所述第一语音与所述第二语音的说话人不同
S203
调用语音处理模型中编码器分别对所述第一语音和第二语音进行编码，分别获得所述第一语音对应的第一特征，以及所述第二语音对应的第二特征；所述第一特征表征与说话人身份无关的文本信息，所述第二特征表征目标说话人音色信息
S205
调用所述语音处理模型中解码器对所述第一特征和所述第二特征进行解码，获得目标语音特征
S207
对所述目标语音特征进行语音重建，获得转换后的目标语音，所述目标语音为由所述第一语音经过音色转换为所述第二语音的语音

图2

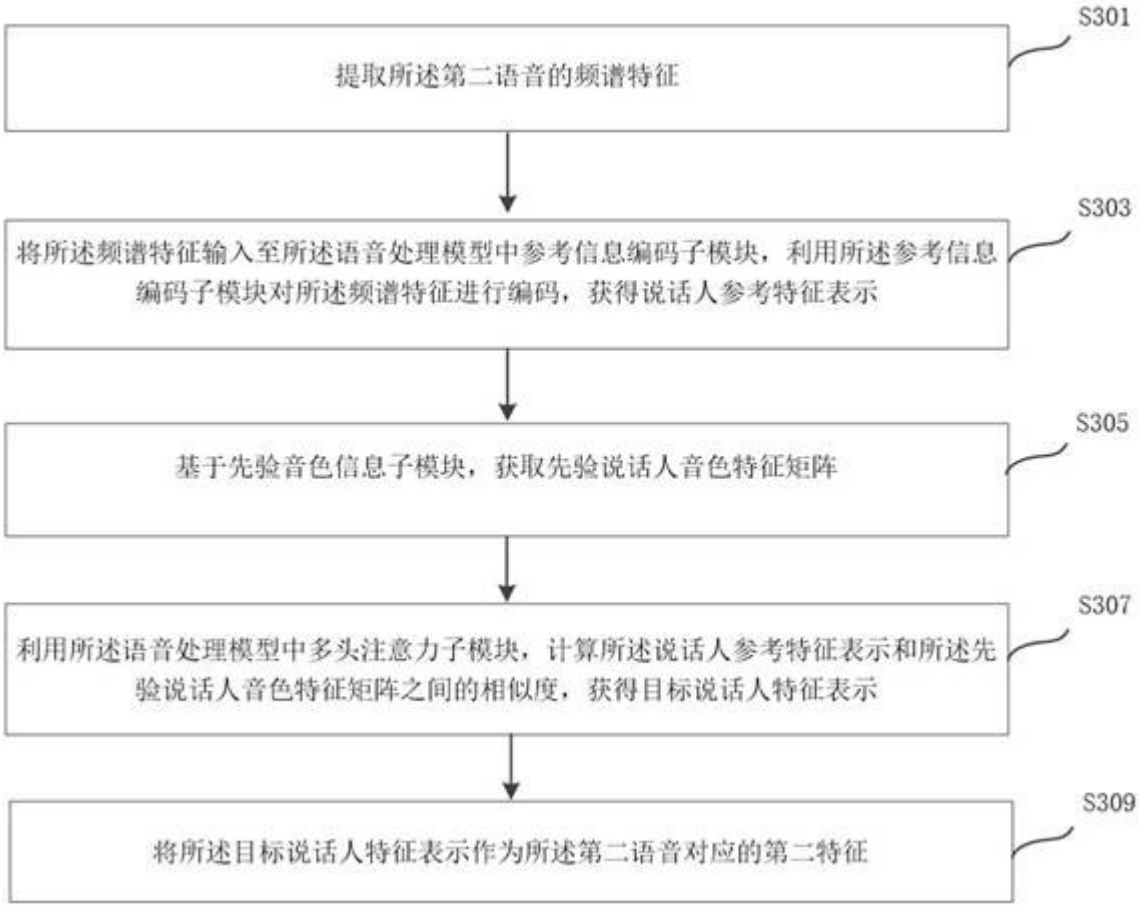

S301
提取所述第二语音的频谱特征
S303
将所述频谱特征输入至所述语音处理模型中参考信息编码子模块，利用所述参考信息编码子模块对所述频谱特征进行编码，获得说话人参考特征表示
S305
基于先验音色信息子模块，获取先验说话人音色特征矩阵
S307
利用所述语音处理模型中多头注意力子模块，计算所述说话人参考特征表示和所述先验说话人音色特征矩阵之间的相似度，获得目标说话人特征表示
S309
将所述目标说话人特征表示作为所述第二语音对应的第二特征

图3

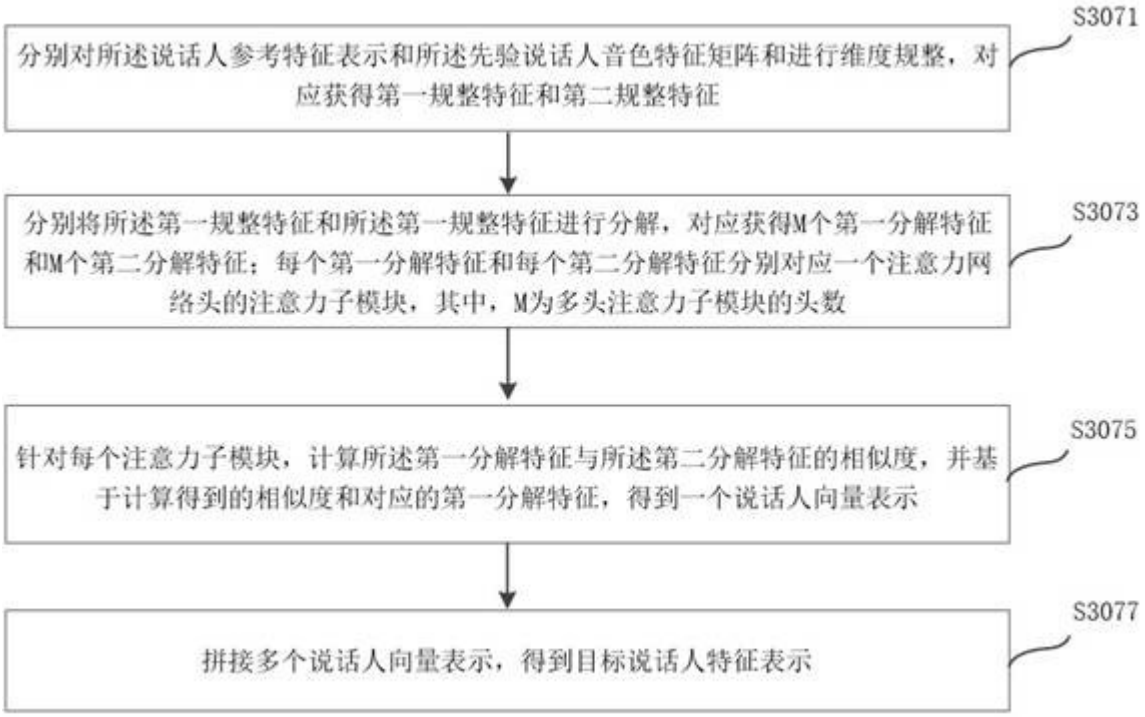

S3071
分别对所述说话人参考特征表示和所述先验说话人音色特征矩阵和进行维度规整，对应获得第一规整特征和第二规整特征
S3073
分别将所述第一规整特征和所述第一规整特征进行分解，对应获得M个第一分解特征和M个第二分解特征；每个第一分解特征和每个第二分解特征分别对应一个注意力网络头的注意力子模块，其中，M为多头注意力子模块的头数
S3075
针对每个注意力子模块，计算所述第一分解特征与所述第二分解特征的相似度，并基于计算得到的相似度和对应的第一分解特征，得到一个说话人向量表示
S3077
拼接多个说话人向量表示，得到目标说话人特征表示

图4

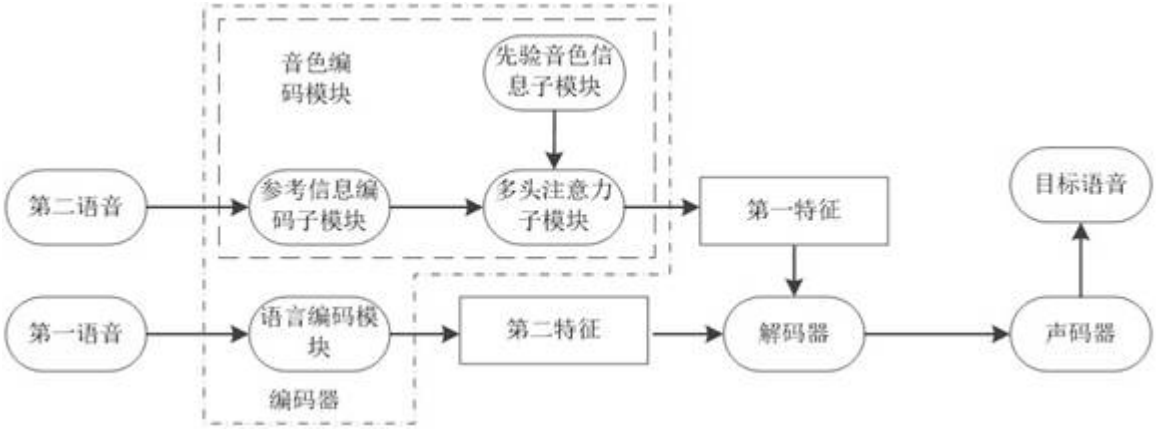

音色编码模块
先验音色信息子模块
第二语音
参考信息编码子模块
多头注意力子模块
第一特征
目标语音
第一语音
语言编码模块
第二特征
解码器
声码器
编码器

图5


语音处理装置
获取模块610
编码模块620
解码模块630
重建模块640

图6

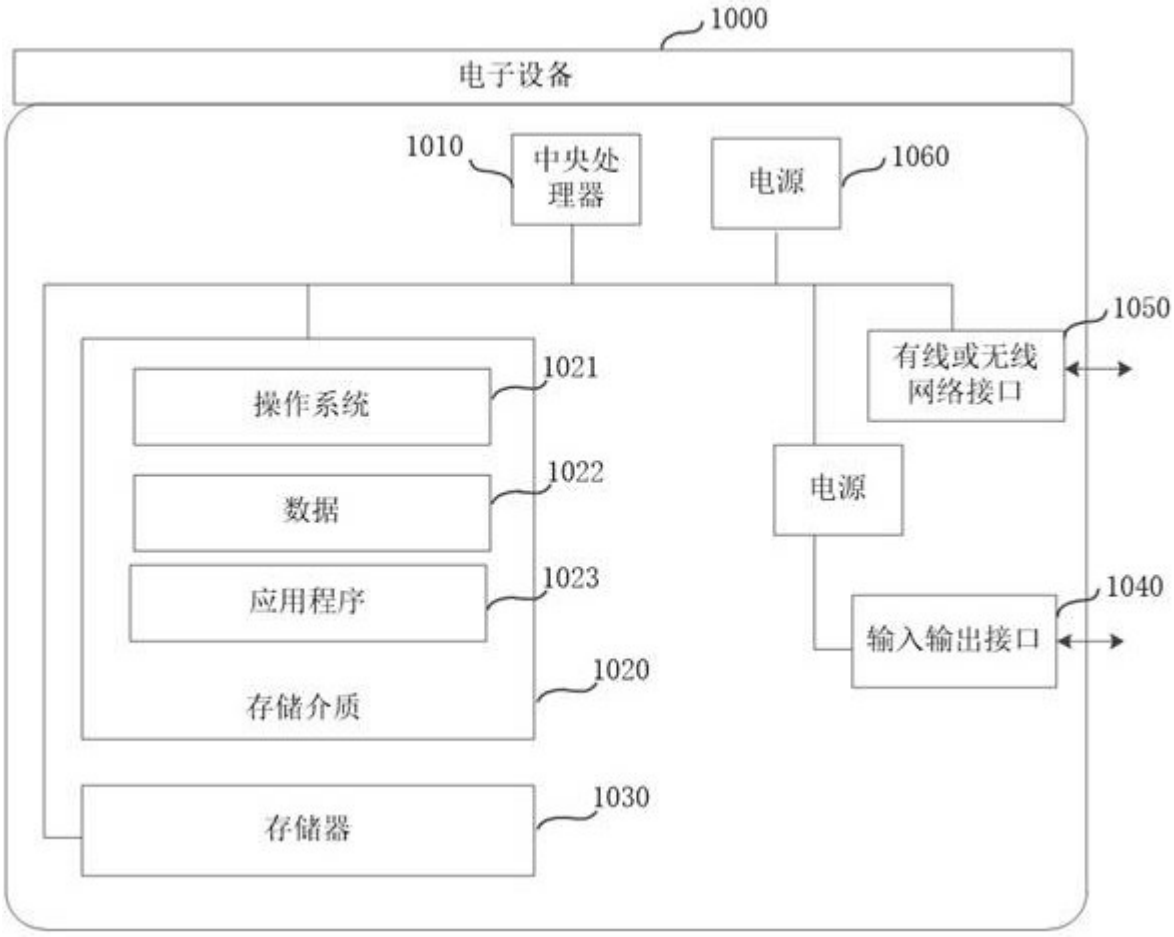

1000
电子设备
1010
中央处理器
电源
1060
1050
有线或无线网络接口
1021
操作系统
1022
数据
电源
1023
应用程序
1040
输入输出接口
1020
存储介质
1030
存储器

图7