(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115273831 A
(43)申请公布日 2022.11.01

(21)申请号 202210916630.8

(22)申请日 2022.08.01

(71)申请人 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 北京清亦华知识产权代理事务所(普通合伙) 11201
专利代理师 孟洋

(51)Int.Cl.
G10L 15/06(2013.01)
G10L 15/01(2013.01)
G10L 13/027(2013.01)
G10L 21/013(2013.01)

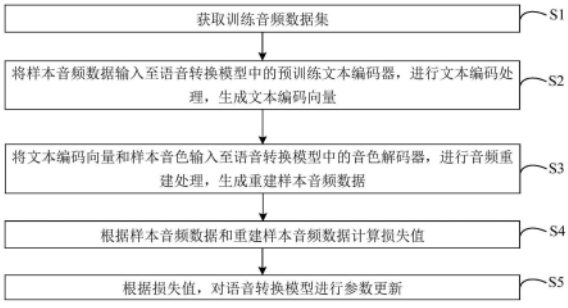
权利要求书2页 说明书15页 附图5页

(54)发明名称

语音转换模型训练方法、语音转换方法和装置

(57)摘要

本公开关于一种语音转换模型训练方法、语音转换方法和装置，涉及计算机技术领域。其中，语音转换模型训练方法，包括：获取训练音频数据集；将样本音频数据输入至语音转换模型中的预训练文本编码器，进行文本编码处理，生成文本编码向量；将文本编码向量和样本音色输入至语音转换模型中的音色解码器，进行音频重建处理，生成重建样本音频数据；根据样本音频数据和重建样本音频数据计算损失值；根据损失值，对语音转换模型进行参数更新。由此，采用联合训练的方式得到的训练好的语音转换模型，能够端到端的进行语音转换，根据文本编码向量进行音频重建处理，具有较好的变声音准以及情感跟随能力。



CN 115273831 A

1.一种语音转换模型训练方法,其特征在于,包括:

获取训练音频数据集;其中,所述训练音频数据集包括:至少一个目标对象的样本音频数据,以及所述样本音频数据对应的样本音色;

将所述样本音频数据输入至语音转换模型中的预训练文本编码器,进行文本编码处理,生成文本编码向量;

将所述文本编码向量和所述样本音色输入至语音转换模型中的音色解码器,进行音频重建处理,生成重建样本音频数据;

根据所述样本音频数据和所述重建样本音频数据计算损失值;

根据所述损失值,对所述语音转换模型进行参数更新。

2.根据权利要求1所述的方法,其特征在于,所述损失值包括第一损失值和/或第二损失值,所述根据所述样本音频数据和所述重建样本音频数据计算损失值,包括:

计算所述样本音频数据的样本音色声学特征和所述重建样本音频数据的重建音色声学特征之间的第一损失值,和/或

将所述重建样本音频数据输入至所述预训练文本编码器,生成重建文本编码向量,计算所述文本编码向量和所述重建文本编码向量之间的第二损失值。

3.根据权利要求2所述的方法,其特征在于,所述根据所述损失值,对所述语音转换模型进行参数更新,包括以下至少一个:

响应于连续预设次数的所述损失值中至少一个大于第一预设值,对所述语音转换模型中的所述音色解码器进行参数更新;响应于连续预设次数的所述损失值的极差大于第二预设值,对所述语音转换模型中的所述音色解码器进行参数更新;

响应于连续预设次数的所述损失值均小于第一预设值,对所述语音转换模型中的所述预训练文本编码器和所述音色解码器进行参数更新;

响应于连续预设次数的所述损失值的极差小于第二预设值,对所述语音转换模型中的所述预训练文本编码器和所述音色解码器进行参数更新。

4.根据权利要求1至3中任一项所述的方法,其特征在于,所述方法,还包括:

获取训练文本数据集;其中,所述训练文本数据集包括:至少一个对象的预训练音频数据和所述预训练音频数据对应的预训练文本数据;

将所述预训练音频数据输入至文本编码器,进行文本编码,生成预训练文本编码向量;

将所述预训练文本编码向量输入至文本解码器,进行解码处理,生成目标文本数据;

根据所述目标文本数据和所述预训练文本数据计算第一预训练损失值;

根据所述第一预训练损失值,对所述文本编码器进行参数更新,以生成所述预训练文本编码器。

5.根据权利要求4所述的方法,其特征在于,所述训练文本数据集还包括所述预训练音频数据对应的预训练单音素数据,其中,还包括:

将所述预训练文本编码向量输入至音素解码器,进行音素解码,生成目标单音素数据;

根据所述目标单音素数据和所述预训练单音素数据计算第二预训练损失值;

根据所述第一预训练损失值和所述第二预训练损失值,对所述文本编码器进行参数更新,以生成所述预训练文本编码器。

6.一种语音转换方法,其特征在于,包括:

获取原始对象的音频数据；

确定需要转换成的目标对象的目标音色；

将所述音频数据和所述目标音色输入至训练好的语音转换模型，进行语音转换处理，生成所述目标对象的目标音频数据；其中，所述训练好的语音转换模型为根据权利要求1至5中任一项所述的方法训练得到的。

7.根据权利要求6所述的方法，其特征在于，所述训练好的语音转换模型包括：训练好的文本编码器和训练好的音色解码器，其中，所述将所述音频数据和所述目标音色输入至训练好的语音转换模型，生成所述目标对象的目标音频数据，包括：

将所述音频数据输入至所述训练好的文本编码器，进行文本编码处理，生成目标文本编码向量；

将所述目标文本编码向量和所述目标音色输入至所述训练好的音色解码器，进行解码处理，生成所述目标对象的所述目标音频数据。

8.一种语音转换模型训练装置，其特征在于，包括：

数据集获取单元，用于获取训练音频数据集；其中，所述训练音频数据集包括：至少一个目标对象的样本音频数据，以及所述样本音频数据对应的样本音色；

编码模块，用于将所述样本音频数据输入至语音转换模型中的预训练文本编码器，进行文本编码处理，生成文本编码向量；

语音重建单元，用于将所述文本编码向量和所述样本音色输入至所述语音转换模型中的音色解码器，进行音频重建处理，生成重建样本音频数据；

损失计算单元，用于根据所述样本音频数据和所述重建样本音频数据计算损失值；

模型更新单元，用于根据所述损失值，对所述语音转换模型进行参数更新。

9.一种语音转换装置，其特征在于，包括：

音频数据获取单元，用于获取原始对象的音频数据；

目标音色确定单元，用于确定需要转换成的目标对象的目标音色；

目标音频获取单元，用于将所述音频数据和所述目标音色输入至训练好的语音转换模型，进行语音转换处理，生成所述目标对象的目标音频数据；其中，所述训练好的语音转换模型为根据权利要求1至5中任一项所述的方法训练得到的。

10.一种电子设备，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5中任一项所述的方法，或所述处理器被配置为执行所述指令，以实现如权利要求6或7所述的方法。

11.一种存储介质，其特征在于，当所述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5中任一项所述的方法，或当所述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6或7所述的方法。

12.一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现根据权利要求1至5中任一项所述的方法，或所述计算机程序在被处理器执行时实现根据权利要求6或7所述的方法。

语音转换模型训练方法、语音转换方法和装置

技术领域

[0001] 本公开涉及计算机技术领域，尤其涉及一种语音转换模型训练方法、语音转换方法和装置。

背景技术

[0002] 随着科学技术的不断发展，人们已经能够通过电子设备(如手机、笔记本电脑、平板电脑、智能家居等)进行声音的录制和播放。

[0003] 相关技术中，在电影配音，短视频变声，虚拟人等方面，存在需要进行语音转换的需求。其中，语音转换是指在保留原始说话人音频数据的语言内容的情况下，将原始说话人音频数据中的音色转移为目标说话人的音色，获取目标说话人的音频数据。

[0004] 其中，语音转换通常使用两个模型，一个模型用于将原始说话人的音频数据转换为文本数据，另一个模型用于将文本数据转换为目标说话人的音频数据。但是，语音转换需要两个模型处理，语音转换的变音音准受限于文本数据的转换，使得变音音准不佳。

发明内容

[0005] 本公开提供一种语音转换模型训练方法、语音转换方法和装置，以采用联合训练的方式得到的训练好的语音转换模型，端到端的进行语音转换。本公开的技术方案如下：

[0006] 根据本公开实施例的第一方面，提供一种语音转换模型训练方法，包括：获取训练音频数据集；其中，所述训练音频数据集包括：至少一个目标对象的样本音频数据，以及所述样本音频数据对应的样本音色；将所述样本音频数据输入至语音转换模型中的预训练文本编码器，进行文本编码处理，生成文本编码向量；将所述文本编码向量和所述样本音色输入至语音转换模型中的音色解码器，进行音频重建处理，生成重建样本音频数据；根据所述样本音频数据和所述重建样本音频数据计算损失值；根据所述损失值，对所述语音转换模型进行参数更新。

[0007] 在一些实施例中，所述损失值包括第一损失值和/或第二损失值，所述根据所述样本音频数据和所述重建样本音频数据计算损失值，包括：计算所述样本音频数据的样本音色声学特征和所述重建样本音频数据的重建音色声学特征之间的第一损失值，和/或将所述重建样本音频数据输入至所述预训练文本编码器，生成重建文本编码向量，计算所述文本编码向量和所述重建文本编码向量之间的第二损失值。

[0008] 在一些实施例中，所述根据所述损失值，对所述语音转换模型进行参数更新，包括以下至少一个：

[0009] 响应于连续预设次数的所述损失值中至少一个大于第一预设值，对所述语音转换模型中的所述音色解码器进行参数更新；

[0010] 响应于连续预设次数的所述损失值的极差大于第二预设值，对所述语音转换模型中的所述音色解码器进行参数更新；

[0011] 响应于连续预设次数的所述损失值均小于第一预设值，对所述语音转换模型中的

所述预训练文本编码器和所述音色解码器进行参数更新；

[0012] 响应于连续预设次数的所述损失值的极差小于第二预设值，对所述语音转换模型中的所述预训练文本编码器和所述音色解码器进行参数更新。

[0013] 在一些实施例中，所述方法，还包括：获取训练文本数据集；其中，所述训练文本数据集包括：至少一个对象的预训练音频数据和所述预训练音频数据对应的预训练文本数据；将所述预训练音频数据输入至文本编码器，进行文本编码，生成预训练文本编码向量；将所述预训练文本编码向量输入至文本解码器，进行解码处理，生成目标文本数据；根据所述目标文本数据和所述预训练文本数据计算预训练损失值；根据所述预训练损失值，对所述文本编码器进行参数更新，以生成所述预训练文本编码器。

[0014] 在一些实施例中，所述训练文本数据集还包括所述预训练音频数据对应的预训练单音素数据，其中，还包括：将所述预训练文本编码向量输入至音素解码器，进行音素解码，生成目标单音素数据；根据所述目标单音素数据和所述预训练单音素数据计算第二预训练损失值；根据所述第一预训练损失值和所述第二预训练损失值，对所述文本编码器进行参数更新，以生成所述预训练文本编码器。

[0015] 根据本公开实施例的第二方面，提供语音转换方法，包括：获取原始对象的音频数据；确定需要转换成的目标对象的目标音色；将所述音频数据和所述目标音色输入至训练好的语音转换模型，进行语音转换处理，生成所述目标对象的目标音频数据；其中，所述训练好的语音转换模型为根据上面一些实施例所述的方法训练得到的。

[0016] 在一些实施例中，所述训练好的语音转换模型包括：训练好的文本编码器和训练好的音色解码器，其中，所述将所述音频数据和所述目标音色输入至训练好的语音转换模型，生成所述目标对象的目标音频数据，包括：将所述音频数据输入至所述训练好的文本编码器，进行文本编码处理，生成目标文本编码向量；将所述目标文本编码向量和所述目标音色输入至所述训练好的音色解码器，进行解码处理，生成所述目标对象的所述目标音频数据。

[0017] 根据本公开实施例的第三方面，提供一种语音转换模型训练装置，包括：数据集获取单元，用于获取训练音频数据集；其中，所述训练音频数据集包括：至少一个目标对象的样本音频数据，以及所述样本音频数据对应的样本音色；编码模块，用于将所述样本音频数据输入至语音转换模型中的预训练文本编码器，进行文本编码处理，生成文本编码向量；语音重建单元，用于将所述文本编码向量和所述样本音色输入至所述语音转换模型中的音色解码器，进行音频重建处理，生成重建样本音频数据；损失计算单元，用于根据所述样本音频数据和所述重建样本音频数据计算损失值；模型更新单元，用于根据所述损失值，对所述语音转换模型进行参数更新。

[0018] 根据本公开实施例的第四方面，提供一种语音转换装置，包括：音频数据获取单元，用于获取原始对象的音频数据；目标音色确定单元，用于确定需要转换成的目标对象的目标音色；目标音频获取单元，用于将所述音频数据和所述目标音色输入至训练好的语音转换模型，进行语音转换处理，生成所述目标对象的目标音频数据；其中，所述训练好的语音转换模型为根据上面一些实施例的方法训练得到的。

[0019] 根据本公开实施例的第五方面，提供一种电子设备，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，以实现如上述第一

方面所述的语音转换模型训练方法，或者所述处理器被配置为执行所述指令，以实现如上述第二方面所述的语音转换方法。

[0020] 根据本公开实施例的第六方面，提供一种存储介质，当所述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第一方面所述的语音转换模型训练方法，或者当所述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第二方面所述的语音转换方法。

[0021] 根据本公开实施例的第器方面，提供一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现如上面第一方面所述的语音转换模型训练方法，或者所述计算机程序在被处理器执行时实现如上面第二方面所述的语音转换方法。

[0022] 本公开的实施例提供的技术方案至少带来以下有益效果：

[0023] 本公开实施例中提供的语音转换模型训练方法，获取训练音频数据集；其中，训练音频数据集包括：至少一个目标对象的样本音频数据，以及样本音频数据对应的样本音色；将样本音频数据输入至语音转换模型中的预训练文本编码器，进行文本编码处理，生成文本编码向量；将文本编码向量和样本音色输入至语音转换模型中的音色解码器，进行音频重建处理，生成重建样本音频数据；根据样本音频数据和重建样本音频数据计算损失值；根据损失值，对语音转换模型进行参数更新。由此，采用联合训练的方式得到的训练好的语音转换模型，能够端到端的进行语音转换，根据文本编码向量进行音频重建处理，具有较好的变声音准以及情感跟随能力。

[0024] 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

[0025] 此处的附图被并入说明书中并构成本说明书的一部分，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。

[0026] 图1为本公开实施例提供的一种语音转换模型训练方法的流程图；

[0027] 图2为本公开实施例提供的计算损失值的方法的流程图；

[0028] 图3为本公开实施例提供的另一种语音转换模型训练方法的流程图；

[0029] 图4为本公开实施例提供的一种语音转换方法的流程图；

[0030] 图5为本公开实施例提供的语音转换方法中S30的流程图；

[0031] 图6为本公开实施例提供的一种语音转换模型训练装置的结构图；

[0032] 图7为本公开实施例提供的另一种语音转换模型训练装置的结构图；

[0033] 图8为本公开实施例提供的另一种语音转换模型训练装置的结构图；

[0034] 图9为本公开实施例提供的一种语音转换装置的结构图；

[0035] 图10为本公开实施例提供的语音转换装置中目标音频获取单元的结构图；

[0036] 图11为本公开实施例提供的一种电子设备的结构图。

具体实施方式

[0037] 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。

[0038] 除非上下文另有要求，否则，在整个说明书和权利要求书中，术语“包括”被解释为开放、包含的意思，即为“包含，但不限于”。在说明书的描述中，术语“一些实施例”等旨在表明与该实施例或示例相关的特定特征、结构、材料或特性包括在本公开的至少一个实施例或示例中。上述术语的示意性表示不一定是指同一实施例或示例。此外，所述的特定特征、结构、材料或特点可以以任何适当方式包括在任何一个或多个实施例或示例中。

[0039] 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。

[0040] 以下，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。在本发明的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。

[0041] 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。

[0042] 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。

[0043] 需要说明的是，本公开实施例的语音转换模型训练方法可以由本公开实施例的语音转换模型训练装置执行，该语音转换模型训练装置可以由软件和/或硬件的方式实现，该语音转换模型训练装置可配置在电子设备中，其中，电子设备可以安装并运行语音转换模型训练程序。其中，电子设备可以包括但不限于智能手机、平板电脑、电脑等具有各种操作系统的硬件设备。

[0044] 需要说明的是，本公开实施例的语音转换方法可以由本公开实施例的语音转换装置执行，该语音转换装置可以由软件和/或硬件的方式实现，该语音转换装置可配置在电子设备中，其中，电子设备可以安装并运行语音转换程序。其中，电子设备可以包括但不限于智能手机、平板电脑、电脑等具有各种操作系统的硬件设备。

[0045] 图1为本公开实施例提供的一种语音转换模型训练方法的流程图。

[0046] 如图1所示，本公开实施例提供的语音转换模型训练方法，包括但不限于以下步骤：

[0047] S1：获取训练音频数据集；其中，训练音频数据集包括：至少一个目标对象的样本音频数据，以及样本音频数据对应的样本音色。

[0048] 本公开实施例中，获取训练音频数据集，其中，训练音频数据集包括：至少一个目标对象的样本音频数据，以及样本音频数据对应的样本音色。

[0049] 其中，目标对象可以为具有特定音色的用户，例如：影视剧演员、名人、动画角色等。当然，目标对象还可以为上述示例外的普通用户，在其具有特定音色，且能够获取音频数据的情况下，也可以作为目标对象，本公开实施例对此不作具体限制。

[0050] 本公开实施例中，获取训练音频数据集，训练音频数据集包括一个或多个目标对象的样本音频数据。其中，一个目标对象的样本音频数据，可以包括一段或多段语音数据。

[0051] 其中，在获取目标对象的样本音频数据的基础上，可以对样本音频数据标记对应的样本音色。本公开实施例中，可以采用统一的标记规则对样本音频数据标记样本音色。

[0052] 示例性的，提取目标对象的样本音频数据的声学特征，以声学特征作为样本音色；或者，对每一个目标对象的样本音频数据进行编号，属于同一目标对象的样本音频数据编号相同，不同目标对象的样本音频数据编号不同，等。

[0053] 本公开实施例中，可以通过专用的音频采集装置，采集目标对象的样本音频数据(例如目标对象为普通用户的情况)，或者，还可以从目标对象公开的音频数据中，获取样本音频数据(例如目标对象为影视剧演员、名人、动画角色的情况)。可以根据目标对象的不同，采用不同的音频采集方式，获取样本音频数据。

[0054] S2：将样本音频数据输入至语音转换模型中的预训练文本编码器，进行文本编码处理，生成文本编码向量。

[0055] 本公开实施例中，语音转换模型包括预训练文本编码器，其中，预训练文本编码器可以为预先经过训练的文本编码器。

[0056] 其中，预训练文本编码器可以根据样本音频数据，生成文本编码向量，提取样本音频数据中的文本信息。

[0057] 本公开实施例中，预训练文本编码器可以为预训练的Conformer-encoder。其中，预训练文本编码器可以为预训练的语音识别模型的分类层之前的部分，预训练的语音识别模型可以为预训练的Conformer-encoder-decoder。

[0058] 需要说明的是，获取预训练的语音识别模型可以采用相关技术中的方法，本公开实施例对此不作具体限制。

[0059] 本公开实施例中，在获取预训练的语音识别模型的情况下，使用预训练的语音识别模型的预训练文本编码器，获取样本音频数据中的文本编码向量。

[0060] 可以理解的是，预训练文本编码器为预训练的语音识别模型中的部分，经过预训练的语音识别模型能够很好的识别样本音频数据中的文本信息，所生成的文本编码向量较为准确。

[0061] S3：将文本编码向量和样本音色输入至语音转换模型中的音色解码器，进行音频重建处理，生成重建样本音频数据。

[0062] 本公开实施例中，在将样本音频数据输入至语音转换模型的预训练文本编码器，生成文本编码向量的情况下，将文本编码向量和样本音色输入至语音转换模型的音色解码器，生成重建样本音频数据。

[0063] 本公开实施例中，根据文本编码向量和样本音色，输入至语音转换模型中的音色解码器，进行音频重建处理，生成重建样本音频数据，相比于相关技术中，先使用一个模型将原始说话人的音频数据转换为文本数据，另一个模型利用转换的文本数据进一步生成目标说话人的音频数据。

[0064] 本公开实施例提供的方法，无需将原始说话人的音频数据转换为文本数据，而是利用样本音频数据输入至语音转换模型的预训练文本编码器，生成的文本编码向量，将文本编码向量和样本音色输入至语音转换模型的音色解码器，生成重建样本音频数据。从而，

能够使得生成的重建样本音频数据，不再受限于文本数据的转换，能够具有很好的变音音准。

[0065] 其中，音色解码器可以为Decoder，可以根据文本编码向量和样本音色，生成重建样本音频数据。重建样本音频数据可以为样本音色的音频数据。

[0066] S4：根据样本音频数据和重建样本音频数据计算损失值。

[0067] 本公开实施例中，在获取重建样本音频数据的情况下，可以根据样本音频数据和重建样本音频数据计算损失值。

[0068] 其中，可以提取样本音频数据的样本音色声学特征，以及提取重建样本音频数据的重建音色声学特征，进而，计算样本音色声学特征和重建音色声学特征之间的损失值。或者，还可以根据样本音频数据和重建样本音频数据的其他参数特征，计算损失值，或者还可以根据样本音频数据和重建样本音频数据进一步处理后的表征，计算损失值等，本公开实施例对此不作具体限制。

[0069] S5：根据损失值，对语音转换模型进行参数更新。

[0070] 本公开实施例中，根据样本音频数据和重建样本音频数据计算损失值，根据损失值，对语音转换模型进行参数更新。

[0071] 需要说明的是，本公开实施例中，根据损失值，对语音转换模型进行参数更新，可以对语音转换模型的预训练文本编码器进行参数更新，或者对语音转换模型的音色解码器进行参数更新，或者同时对语音转换模型的预训练文本编码器和音色编码器进行参数更新。

[0072] 其中，在根据多个目标对象的样本音频数据和样本音频数据对应的样本音色，计算损失值，损失值足够小且稳定的情况下，可以判断此时语音转换模型已达到需求，可以得到训练好的语音转换模型。

[0073] 本公开实施例中，将样本音频数据输入至语音转换模型的预训练文本编码器，生成文本编码向量，将文本编码向量和样本音色输入至语音转换模型的音色解码器，生成重建样本音频数据，采用联合训练的方式，各个模块均能贡献信息，能够提升语音转换的性能。并且，经过联合训练得到的训练好的语音转换模型为端到端的语音转换系统，具有较好的变声音准以及情感跟随能力。

[0074] 通过实施本公开实施例，获取训练音频数据集；其中，训练音频数据集包括：至少一个目标对象的样本音频数据，以及样本音频数据对应的样本音色；将样本音频数据输入至语音转换模型的预训练文本编码器，生成文本编码向量；将文本编码向量和样本音色输入至语音转换模型的音色解码器，生成重建样本音频数据；根据样本音频数据和重建样本音频数据计算损失值；根据损失值，对语音转换模型进行参数更新。由此，采用联合训练的方式得到的训练好的语音转换模型，能够端到端的进行语音转换，根据文本编码向量进行音频重建处理，具有较好的变声音准以及情感跟随能力。

[0075] 在一些实施例中，损失值包括第一损失值和/或第二损失值，根据样本音频数据和重建样本音频数据计算损失值，包括：计算样本音频数据的样本音色声学特征和重建样本音频数据的重建音色声学特征之间的第一损失值，和/或，将重建样本音频数据输入至预训练文本编码器，生成重建文本编码向量，计算文本编码向量和重建文本编码向量之间的第二损失值。

[0076] 本公开实施例中，根据样本音频数据和重建样本音频数据计算损失值，可以提取样本音频数据的样本音色声学特征，以及提取重建样本音频数据的重建音色声学特征，进而，计算样本音色声学特征和重建音色声学特征之间的第一损失值。

[0077] 本公开实施例中，根据样本音频数据和重建样本音频数据计算损失值，可以将重建样本音频数据输入至预训练文本编码器，生成重建文本编码向量，计算文本编码向量和重建文本编码向量之间的第二损失值。

[0078] 示例性地，如图2所示，本公开实施例中，根据样本音频数据和重建样本音频数据计算损失值，损失值包括第一损失值和第二损失值，可以提取样本音频数据的样本音色声学特征，以及提取重建样本音频数据的重建音色声学特征，进而，计算样本音色声学特征和重建音色声学特征之间的第一损失值；以及将重建样本音频数据输入至预训练文本编码器，生成重建文本编码向量，计算文本编码向量和重建文本编码向量之间的第二损失值。

[0079] 在一些实施例中，计算样本音频数据的样本音色声学特征和重建样本音频数据的重建音色声学特征之间的第一损失值，包括：

[0080] 获取第i帧样本音色声学特征Y_i和重建音色声学特征Y_i^{pred}；其中，i为正整数；

[0081] 根据第i帧样本音色声学特征Y_i和重建音色声学特征Y_i^{pred}，计算第一损失$L_{reconst}$；其中，第一损失值$L_{reconst}$满足如下关系：

[0082] $L_{reconst}=\frac{1}{N}\sum_{i=1}^{N}\left|Y_i-Y_i^{pred}\right|$；其中，N为正整数。

[0083] 本公开实施例中，N可以为样本音频数据的总帧数，通过计算每一帧样本音色声学特征Y_i和重建音色声学特征Y_i^{pred}的差的绝对值，并求和求平均，得到第一损失值$L_{reconst}$。

[0084] 在一些实施例中，根据损失值，对语音转换模型进行参数更新，包括以下至少一个：

[0085] 响应于连续预设次数的损失值中至少一个大于第一预设值，对语音转换模型中的音色解码器进行参数更新；

[0086] 响应于连续预设次数的损失值的极差大于第二预设值，对语音转换模型中的音色解码器进行参数更新；

[0087] 响应于连续预设次数的损失值均小于第一预设值，对语音转换模型中的预训练文本编码器和音色解码器进行参数更新；

[0088] 响应于连续预设次数的损失值的极差小于第二预设值，对语音转换模型中的预训练文本编码器和音色解码器进行参数更新。

[0089] 本公开实施例中，根据损失值，对语音转换模型进行参数更新，包括：在计算样本音频数据的样本音色声学特征和重建样本音频数据的重建音色声学特征之间的第一损失值，且连续预设次数的损失值中至少一个大于第一预设值的情况下，对语音转换模型中的音色解码器进行参数更新。

[0090] 其中，预设次数可以为50次、或100次等，第一预设值可以根据需要进行设置，例如，0.2、0.3等，本公开实施例对此不作具体限制。

[0091] 本公开实施例中，根据损失值，对语音转换模型进行参数更新，包括：在计算样本音频数据的样本音色声学特征和重建样本音频数据的重建音色声学特征之间的第一损失

值，且连续预设次数的损失值均小于第一预设值的情况下，对语音转换模型中的预训练文本编码器和音色解码器进行参数更新。

[0092] 本公开实施例中，根据损失值，对语音转换模型进行参数更新，包括：在将重建样本音频数据输入至预训练文本编码器，生成重建文本编码向量，计算文本编码向量和重建文本编码向量之间的第二损失值，且连续预设次数的损失值的极差大于第二预设值的情况下，对语音转换模型中的音色解码器进行参数更新。

[0093] 其中，预设次数可以为50次、或100次等，第二预设值可以根据需要进行设置，例如，0.05、0.01等，本公开实施例对此不作具体限制。

[0094] 本公开实施例中，根据损失值，对语音转换模型进行参数更新，包括：在将重建样本音频数据输入至预训练文本编码器，生成重建文本编码向量，计算文本编码向量和重建文本编码向量之间的第二损失值，且连续预设次数的损失值的极差小于第二预设值的情况下，对语音转换模型中的预训练文本编码器和音色解码器进行参数更新。

[0095] 可以理解的是，因为语音转换模型的预训练文本编码器Conformer-encoder中的参数已经在语音识别任务上收敛到稳定的状态，基于特定人的语音转换仅需要进行小规模的更新，而语音转换模型的音色解码器的参数是随机更新的，没有先验信息参考，因此对联合训练时两个模块的参数优化设计不同的策略。

[0096] 在前若干次训练，语音转换模型的预训练文本编码器Conformer-encoder模块参数不更新，仅更新音色解码器的参数；然后，打开预训练文本编码器Conformer-encoder模块的梯度反传，但设计较小的学习率更新。

[0097] 其中，可以在连续预设次数的损失值中至少一个大于第一预设值，对语音转换模型中的音色解码器进行参数更新，之后，在连续预设次数的损失值均小于第一预设值，对语音转换模型中的预训练文本编码器和音色解码器进行参数更新；和/或，在连续预设次数的损失值的极差大于第二预设值，对语音转换模型中的音色解码器进行参数更新，之后，在连续预设次数的损失值的极差小于第二预设值，对语音转换模型中的预训练文本编码器和音色解码器进行参数更新。

[0098] 如图3所示，在一些实施例中，本公开实施例提供的语音转换模型训练方法，还包括：

[0099] S11：获取训练文本数据集；其中，训练文本数据集包括：至少一个对象的预训练音频数据和预训练音频数据对应的预训练文本数据。

[0100] 本公开实施例中，获取训练文本数据集；其中，训练文本数据集包括：至少一个对象的预训练音频数据和预训练音频数据对应的预训练文本数据。

[0101] 其中，训练文本数据集中，包括一个或多个对象的预训练音频数据，该对象可以为：普通用户，或者具有特定音色的用户。可以理解的是，普通用户可以包括老年人、中年人、年轻人、儿童等。

[0102] 本公开实施例中，获取一个或多个对象的预训练音频数据，以及预训练音频数据对应的训练文本数据，此过程可以采用人工标记，或者还可以借助相关技术中较为成熟的技术实现，本公开实施例对此不作具体限制。

[0103] 其中，一个对象的预训练音频数据，可以包括一段或多段语音数据。

[0104] S12：将预训练音频数据输入至文本编码器，进行文本编码，生成预训练文本编码

向量。

[0105] 本公开实施例中，在获取一个或多个对象的预训练音频数据的情况下，将预训练音频数据输入至文本编码器，生成预训练文本编码向量。其中，文本编码器可以为Conformer-encoder。

[0106] 本公开实施例中，将预训练音频数据输入至Conformer-encoder，生成预训练文本编码向量。

[0107] S13：将预训练文本编码向量输入至文本解码器，进行解码处理，生成目标文本数据。

[0108] 本公开实施例中，在将预训练音频数据输入至文本编码器，生成预训练文本编码向量的情况下，将预训练文本编码向量进一步输入至文本解码器，生成目标文本数据。其中，文本解码器可以为Decoder，能够根据预训练文本编码向量，解码生成目标文本数据。

[0109] S14：根据目标文本数据和预训练文本数据计算第一预训练损失值。

[0110] 本公开实施例中，在将预训练音频数据输入至文本编码器，生成预训练文本编码向量，进一步将预训练文本编码向量输入至文本解码器，生成目标文本数据的情况下，得到目标文本数据。

[0111] 在此基础上，可以根据目标文本数据和预训练文本数据计算第一预训练损失值。

[0112] 可以理解的是，目标文本数据和预训练文本数据均为文本，可以比较文本的区别计算第一预训练损失值，或者还可以进一步获取目标文本数据和预训练文本数据对应的文本向量，计算第一预训练损失值，或者还可以采用相关技术中的方法等，本公开实施例对此不作具体限制。

[0113] S15：根据第一预训练损失值，对文本编码器进行参数更新，以获取预训练文本编码器。

[0114] 本公开实施例中，根据目标文本数据和预训练文本数据计算第一预训练损失值后，根据第一预训练损失值，对文本编码器进行参数更新，以获取预训练文本编码器。其中，还可以根据第一预训练损失值对文本编码器和文本解码器同时进行参数更新。

[0115] 其中，在根据多个对象的目标文本数据和预训练文本数据计算第一预训练损失值，第一预训练损失值足够小且稳定的情况下，可以判断此时文本编码器和文本解码器已达到需求，可以得到训练好的文本编码器和训练好的文本解码器，生成预训练文本编码器。

[0116] 在一些实施例中，训练文本数据集还包括预训练音频数据对应的预训练单音素数据，其中，还包括：将预训练文本编码向量输入至音素解码器，进行音素解码，生成目标单音素数据；根据目标单音素数据和预训练单音素数据计算第二预训练损失值；根据第一预训练损失值和第二预训练损失值，对文本编码器进行参数更新，以生成预训练文本编码器。

[0117] 本公开实施例中，训练文本数据集还包括预训练音频数据对应的预训练单音素数据，例如：汉语拼音中的a，o，e等。

[0118] 本公开实施例中，将预训练音频数据输入至文本编码器Conformer-encoder，生成预训练文本编码向量之后，在文本编码器Conformer-encoder后添加若干线性层，将预训练文本编码向量输出映射为单音素，生成目标单音素数据。

[0119] 基于此，在获取目标单音素数据和预训练单音素数据的情况下，可以根据目标单音素数据和预训练单音素数据计算第二预训练损失值，使用单音素损失函数作为优化目

标，计算第二预训练损失值。

[0120] 其中，在根据第一预训练损失值，对文本编码器进行参数更新的同时，可以根据第二预训练损失值，对文本编码器进行参数更新，以获取预训练文本编码器。

[0121] 其中，还可以根据第一预训练损失值和第二预训练损失值对文本编码器和文本解码器同时进行参数更新。其中，在第一预训练损失值和第二预训练损失值足够小且稳定的情况下，可以判断此时文本编码器和文本解码器已达到需求，可以得到训练好的文本编码器和训练好的文本解码器，生成预训练文本编码器。

[0122] 图4为本公开实施例提供的一种语音转换方法的流程图。

[0123] 如图4所示，本公开实施例提供的语音转换方法，包括但不限于以下步骤：

[0124] S10：获取原始对象的音频数据。

[0125] 本公开实施例中，获取原始对象的音频数据，原始对象的音频数据可以为用户上传的一段录音数据，或者，用户实时录制的一段录音数据。

[0126] 可以理解的是，本公开实施例提供的语音转换方法，可以由终端设备执行，例如：智能手机，电脑等。用户可以使用终端设备录制一段自己说话的声音，从而，终端设备获取原始对象的音频数据。

[0127] S20：确定需要转换成的目标对象的目标音色。

[0128] 本公开实施例中，可以提供目标对象的目标音色的参照表，从而用户可以根据参照表选择目标对象的目标音色；或者，还可以仅提供目标对象，在用户选择目标对象后，可以根据选择的目标对象，确定目标音色。

[0129] 可以理解的是，用户使用终端设备录制一段自己说话的声音之后，可以选择想要转换成的目标对象的目标音色。

[0130] S30：将音频数据和目标音色输入至训练好的语音转换模型，进行语音转换处理，生成目标对象的目标音频数据；其中，训练好的语音转换模型为根据上面一些实施例的方法训练得到的。

[0131] 本公开实施例中，在获取原始对象的音频数据，以及确定需要转换成的目标对象的目标音色的情况下，将音频数据和目标音色输入至训练好的语音转换模型，可以生成目标对象的目标音频数据。从而实现语音转换。

[0132] 其中，训练好的语音转换模型为根据上面一些实施例的方法进行训练得到的，具体可以参见上述实施例中的相关描述，此处不再赘述。

[0133] 为方便理解，本公开实施例提供一示例性实施例。

[0134] 本公开实施例中，终端设备能够执行语音转换方法，用户可以使用终端设备录制一段自己说话的语音，选择目标对象，例如：选择蜡笔小新，基于此，可以确定蜡笔小新对应的目标音色。

[0135] 之后，将用户录制的语音，和蜡笔小新对应的目标音色输入至训练好的语音转换模型，可以生成蜡笔小新的目标音频数据。例如：用户录制的语音为“你好，我是蜡笔小新”，可以理解的是，用户录制的语音为用户的音色，不同于蜡笔小新的音色。用户录制完成后，可以选择语音转换为蜡笔小新的音色，输入至训练好的语音转换模型，能够生成蜡笔小新的音色的音频数据，所得到的目标音频数据“你好，我是蜡笔小新”的音色为蜡笔小新的音色。

[0136] 需要说明的是，用户也可以先选择目标对象的目标音色，再录制语音，步骤的先后顺序可以根据需要进行调整。

[0137] 通过实施本公开实施例中，获取原始对象的音频数据；确定需要转换成的目标对象的目标音色；将音频数据和目标音色输入至训练好的语音转换模型，生成目标对象的目标音频数据；其中，训练好的语音转换模型为根据上面一些实施例的方法训练得到的。由此，能够实现语音转换，且具有较好的变声音准以及情感跟随能力。

[0138] 如图5所示，在一些实施例中，训练好的语音转换模型包括：训练好的文本编码器和训练好的音色解码器，其中，S30：将音频数据和目标音色输入至训练好的语音转换模型，生成目标对象的目标音频数据，包括：

[0139] S301：将音频数据输入至训练好的文本编码器，进行文本编码处理，生成目标文本编码向量。

[0140] S302：将目标文本编码向量和目标音色输入至训练好的音色解码器，进行解码处理，生成目标对象的目标音频数据。

[0141] 本公开实施例中，训练好的语音转换模型包括：训练好的文本编码器和训练好的音色解码器。其中，训练好的文本编码器可以为训练好的Conformer-encoder，训练好的音色解码器可以为训练好的Decoder。

[0142] 本公开实施例中，将音频数据输入至训练好的文本编码器，生成目标文本编码向量，将目标文本编码向量和目标音色输入至训练好的音色解码器，生成目标对象的目标音频数据。

[0143] 图6为本公开实施例提供的一种语音转换模型训练装置的结构图。

[0144] 如图6所示，该语音转换模型训练装置1，包括：数据集获取单元11、编码模块12、语音重建单元13、损失计算单元14和模型更新单元15。

[0145] 数据集获取单元11，用于获取训练音频数据集；其中，训练音频数据集包括：至少一个目标对象的样本音频数据，以及样本音频数据对应的样本音色。

[0146] 编码模块12，用于将样本音频数据输入至语音转换模型中的预训练文本编码器，进行文本编码处理，生成文本编码向量。

[0147] 语音重建单元13，用于将文本编码向量和样本音色输入至语音转换模型中的音色解码器，进行音频重建处理，生成重建样本音频数据。

[0148] 损失计算单元14，用于根据样本音频数据和重建样本音频数据计算损失值。

[0149] 模型更新单元15，用于根据损失值，对语音转换模型进行参数更新。

[0150] 通过实施本公开实施例，数据集获取单元11获取训练音频数据集；其中，训练音频数据集包括：至少一个目标对象的样本音频数据，以及样本音频数据对应的样本音色；编码模块12将样本音频数据输入至语音转换模型中的预训练文本编码器，进行文本编码处理，生成文本编码向量；语音重建单元13将文本编码向量和样本音色输入至语音转换模型中的音色解码器，进行音频重建处理，生成重建样本音频数据；损失计算单元14根据样本音频数据和重建样本音频数据计算损失值；模型更新单元15根据损失值，对语音转换模型进行参数更新。由此，采用联合训练的方式得到的训练好的语音转换模型，能够端到端的进行语音转换，根据文本编码向量进行音频重建处理，具有较好的变声音准以及情感跟随能力。

[0151] 在一些实施例中，损失值包括第一损失值和/或第二损失值，损失计算单元14，具

体用于:计算样本音频数据的样本音色声学特征和重建样本音频数据的重建音色声学特征之间的第一损失值,和/或将重建样本音频数据输入至预训练文本编码器,生成重建文本编码向量,计算文本编码向量和重建文本编码向量之间的第二损失值。

[0152] 在一些实施例中,模型更新单元15,具体用于:

[0153] 响应于连续预设次数的损失值中至少一个不小于第一预设值,对语音转换模型中的音色解码器进行参数更新;

[0154] 响应于连续预设次数的损失值的极差不小于第二预设值,对语音转换模型中的音色解码器进行参数更新;

[0155] 响应于连续预设次数的损失值均小于第一预设值,对语音转换模型中的预训练文本编码器和音色解码器进行参数更新;

[0156] 响应于连续预设次数的损失值的极差小于第二预设值,对语音转换模型中的预训练文本编码器和音色解码器进行参数更新。

[0157] 如图7所示,在一些实施例中,该语音转换模型训练装置1,还包括:

[0158] 训练文本获取单元61,用于获取训练文本数据集;其中,训练文本数据集包括:至少一个对象的预训练音频数据和预训练音频数据对应的预训练文本数据。

[0159] 文本编码单元62,用于将预训练音频数据输入至文本编码器,进行文本编码,生成预训练文本编码向量。

[0160] 文本解码单元63,用于将预训练文本编码向量输入至文本解码器,进行解码处理,生成目标文本数据。

[0161] 第一训练损失计算单元64,用于根据目标文本数据和预训练文本数据计算第一预训练损失值。

[0162] 第一编码器更新单元65,用于根据第一预训练损失值,对文本编码器进行参数更新,以获取预训练文本编码器。

[0163] 如图8所示,在一些实施例中,训练文本数据集还包括预训练音频数据对应的预训练单音素数据,该语音转换模型训练装置1,还包括:

[0164] 音素解码单元71,用于将预训练文本编码向量输入至音素解码器,进行音素解码,生成目标单音素数据。

[0165] 第二训练损失计算单元72,用于根据目标单音素数据和预训练单音素数据计算第二预训练损失值。

[0166] 第二编码器更新单元73,用于根据第一预训练损失值和第二预训练损失值,对文本编码器进行参数更新,以生成预训练文本编码器。

[0167] 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。

[0168] 本公开实施例中语音转换模型训练装置所能取得的有益效果与上述示例语音转换模型训练方法所能取得的有益效果相同,此处不再赘述。

[0169] 图9为本公开实施例提供的一种语音转换装置的结构图。

[0170] 如图9所示,该语音转换装置80,包括:音频数据获取单元801、目标音色确定单元802和目标音频获取单元803。

[0171] 音频数据获取单元801,用于获取原始对象的音频数据。

[0172] 目标音色确定单元802,用于确定需要转换成的目标对象的目标音色。

[0173] 目标音频获取单元803,用于将音频数据和目标音色输入至训练好的语音转换模型,进行语音转换处理,生成目标对象的目标音频数据;其中,训练好的语音转换模型为根据上面一些实施例的方法训练得到的。

[0174] 通过实施本公开实施例中,音频数据获取单元801获取原始对象的音频数据;目标音色确定单元802确定需要转换成的目标对象的目标音色;目标音频获取单元803将音频数据和目标音色输入至训练好的语音转换模型,进行语音转换处理,生成目标对象的目标音频数据;其中,训练好的语音转换模型为根据上面一些实施例的方法训练得到的。由此,能够实现语音转换,且具有较好的变声音准以及情感跟随能力。

[0175] 如图10所示,在一些实施例中,训练好的语音转换模型,包括:训练好的文本编码器和训练好的音色解码器,其中,目标音频获取单元803,包括:目标文本编码模块8031和目标音频解码模块8032。

[0176] 目标文本编码模块8031,用于将音频数据输入至训练好的文本编码器,进行文本编码处理,生成目标文本编码向量。

[0177] 目标音频解码模块8032,用于将目标文本编码向量和目标音色输入至训练好的音色解码器,进行解码处理,生成目标对象的目标音频数据。

[0178] 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。

[0179] 本公开实施例中语音转换装置所能取得的有益效果与上述示例语音转换方法所能取得的有益效果相同,此处不再赘述。

[0180] 图11为本公开实施例提供的一种用于执行语音转换模型训练方法,或者语音转换方法的电子设备100的框图。

[0181] 示例性地,电子设备100可以是移动电话,计算机,数字广播终端,消息收发设备,游戏控制台,平板设备,医疗设备,健身设备,个人数字助理等。

[0182] 如图11所示,电子设备100可以包括以下一个或多个组件:处理组件101,存储器102,电源组件103,多媒体组件104,音频组件105,输入/输出(I/O)的接口106,传感器组件107,以及通信组件108。

[0183] 处理组件101通常控制电子设备100的整体操作,诸如与显示,电话呼叫,数据通信,相机操作和记录操作相关联的操作。处理组件101可以包括一个或多个处理器1011来执行指令,以完成上述的方法的全部或部分步骤。此外,处理组件101可以包括一个或多个模块,便于处理组件101和其他组件之间的交互。例如,处理组件101可以包括多媒体模块,以方便多媒体组件104和处理组件101之间的交互。

[0184] 存储器102被配置为存储各种类型的数据以支持在电子设备100的操作。这些数据的示例包括用于在电子设备100上操作的任何应用程序或方法的指令,联系人数据,电话簿数据,消息,图片,视频等。存储器102可以由任何类型的易失性或非易失性存储设备或者它们的组合实现,如SRAM(Static Random-Access Memory,静态随机存取存储器),EEPROM(Electrically Erasable Programmable read only memory,带电可擦可编程只读存储器),EPROM(Erasable Programmable Read-Only Memory,可擦除可编程只读存储器),PROM(Programmable read-only memory,可编程只读存储器),ROM(Read-Only Memory,只读存

储器),磁存储器,快闪存储器,磁盘或光盘。

[0185] 电源组件103为电子设备100的各种组件提供电力。电源组件103可以包括电源管理系统,一个或多个电源,及其他与为电子设备100生成、管理和分配电力相关联的组件。

[0186] 多媒体组件104包括在所述电子设备100和用户之间的提供一个输出接口的触控显示屏。在一些实施例中,触控显示屏可以包括LCD(Liquid Crystal Display,液晶显示器)和TP(Touch Panel,触摸面板)。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界,而且还检测与所述触摸或滑动操作相关的持续时间和压力。在一些实施例中,多媒体组件104包括一个前置摄像头和/或后置摄像头。当电子设备100处于操作模式,如拍摄模式或视频模式时,前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。

[0187] 音频组件105被配置为输出和/或输入音频信号。例如,音频组件105包括一个MIC(Microphone,麦克风),当电子设备100处于操作模式,如呼叫模式、记录模式和语音识别模式时,麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器102或经由通信组件108发送。在一些实施例中,音频组件105还包括一个扬声器,用于输出音频信号。

[0188] I/O接口2112为处理组件101和外围接口模块之间提供接口,上述外围接口模块可以是键盘,点击轮,按钮等。这些按钮可包括但不限于:主页按钮、音量按钮、启动按钮和锁定按钮。

[0189] 传感器组件107包括一个或多个传感器,用于为电子设备100提供各个方面的状态评估。例如,传感器组件107可以检测到电子设备100的打开/关闭状态,组件的相对定位,例如所述组件为电子设备100的显示器和小键盘,传感器组件107还可以检测电子设备100或电子设备100一个组件的位置改变,用户与电子设备100接触的存在或不存在,电子设备100方位或加速/减速和电子设备100的温度变化。传感器组件107可以包括接近传感器,被配置用来在没有任何的物理接触时检测附近物体的存在。传感器组件107还可以包括光传感器,如CMOS(Complementary Metal Oxide Semiconductor,互补金属氧化物半导体)或CCD(Charge-coupled Device,电荷耦合元件)图像传感器,用于在成像应用中使用。在一些实施例中,该传感器组件107还可以包括加速度传感器,陀螺仪传感器,磁传感器,压力传感器或温度传感器。

[0190] 通信组件108被配置为便于电子设备100和其他设备之间有线或无线方式的通信。电子设备100可以接入基于通信标准的无线网络,如WiFi,2G或3G,或它们的组合。在一个示例性实施例中,通信组件108经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中,所述通信组件108还包括NFC(Near Field Communication,近场通信)模块,以促进短程通信。例如,在NFC模块可基于RFID(Radio Frequency Identification,射频识别)技术,IrDA(Infrared Data Association,红外数据协会)技术,UWB(Ultra Wide Band,超宽带)技术,BT(Bluetooth,蓝牙)技术和其他技术来实现。

[0191] 在示例性实施例中,电子设备100可以被一个或多个ASIC(Application Specific Integrated Circuit,专用集成电路)、DSP(Digital Signal Processor,数字信号处理

器)、数字信号处理设备(DSPD)、PLD(Programmable Logic Device,可编程逻辑器件)、FPGA(Field Programmable Gate Array,现场可编程逻辑门阵列)、控制器、微控制器、微处理器或其他电子元件实现,用于执行上述语音转换模型训练方法,或者语音转换方法。需要说明的是,本实施例的电子设备的实施过程和技术原理参见前述对本公开实施例的语音转换模型训练方法,或者语音转换方法的解释说明,此处不再赘述。

[0192] 本公开实施例提供的电子设备,可以执行如上面一些实施例所述的语音转换模型训练方法,或者语音转换方法,其有益效果与上述的语音转换模型训练方法,或者语音转换方法的有益效果相同,此处不再赘述。

[0193] 为了实现上述实施例,本公开还提出一种存储介质。

[0194] 其中,该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前所述的语音转换模型训练方法,或者语音转换方法。例如,所述存储介质可以是ROM(Read Only Memory Image,只读存储器)、RAM(Random Access Memory,随机存取存储器)、CD-ROM(Compact Disc Read-Only Memory,紧凑型光盘只读存储器)、磁带、软盘和光数据存储设备等。

[0195] 为了实现上述实施例,本公开还提供一种计算机程序产品,该计算机程序由电子设备的处理器执行时,使得电子设备能够执行如前所述的语音转换模型训练方法,或者语音转换方法。

[0196] 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。本公开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。

[0197] 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。

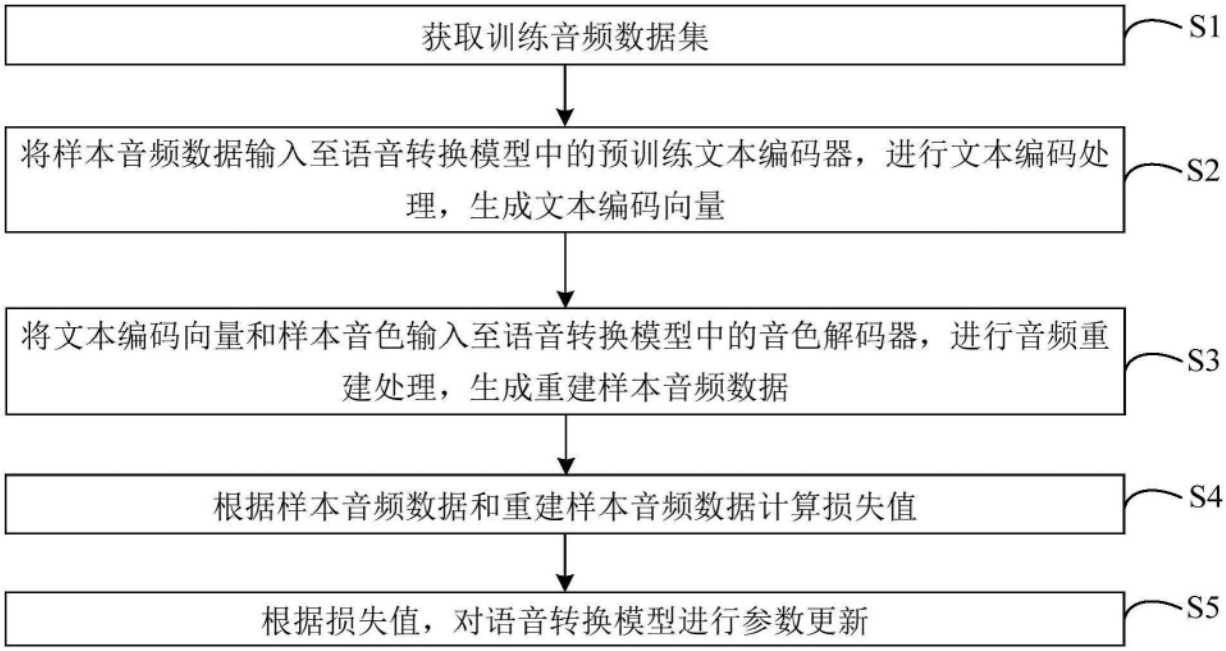

获取训练音频数据集
S1
将样本音频数据输入至语音转换模型中的预训练文本编码器，进行文本编码处理，生成文本编码向量
S2
将文本编码向量和样本音色输入至语音转换模型中的音色解码器，进行音频重建处理，生成重建样本音频数据
S3
根据样本音频数据和重建样本音频数据计算损失值
S4
根据损失值，对语音转换模型进行参数更新
S5

图1

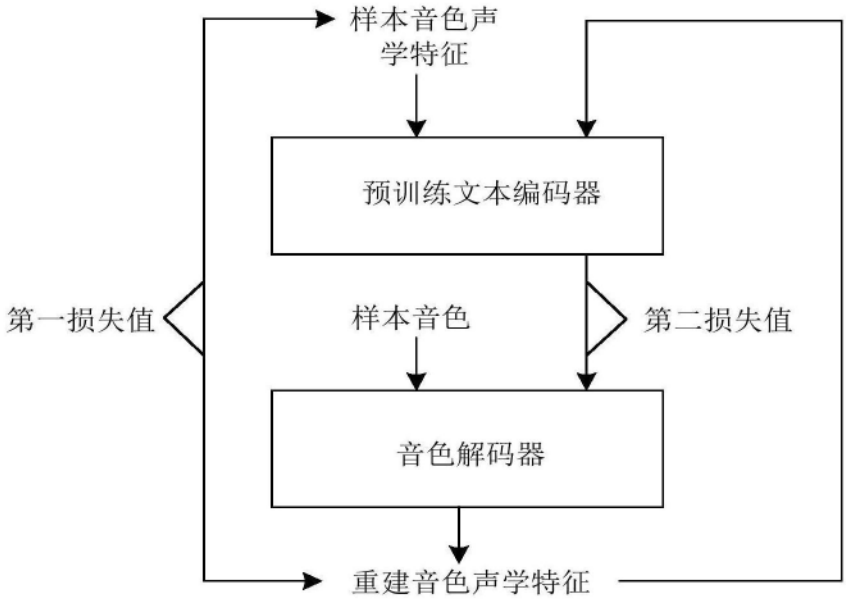

样本音色声学特征
预训练文本编码器
第一损失值
样本音色
第二损失值
音色解码器
重建音色声学特征

图2

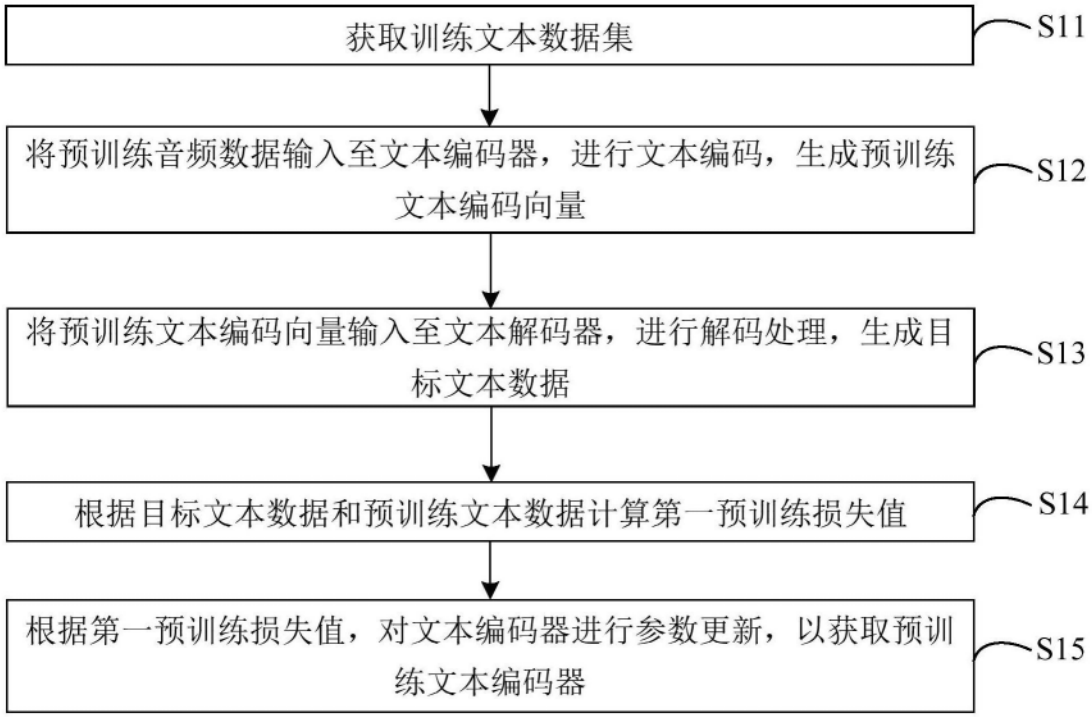

获取训练文本数据集
S11
将预训练音频数据输入至文本编码器，进行文本编码，生成预训练文本编码向量
S12
将预训练文本编码向量输入至文本解码器，进行解码处理，生成目标文本数据
S13
根据目标文本数据和预训练文本数据计算第一预训练损失值
S14
根据第一预训练损失值，对文本编码器进行参数更新，以获取预训练文本编码器
S15

图3

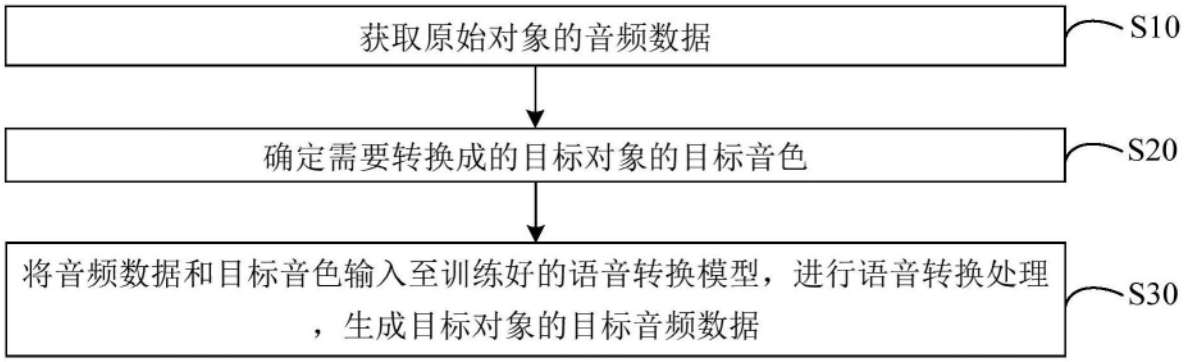

获取原始对象的音频数据
S10
确定需要转换成的目标对象的目标音色
S20
将音频数据和目标音色输入至训练好的语音转换模型，进行语音转换处理，生成目标对象的目标音频数据
S30

图4

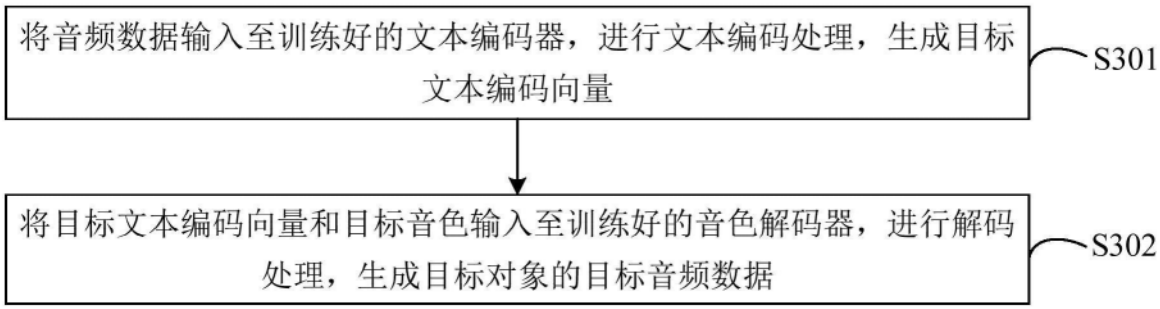

将音频数据输入至训练好的文本编码器，进行文本编码处理，生成目标文本编码向量
S301
将目标文本编码向量和目标音色输入至训练好的音色解码器，进行解码处理，生成目标对象的目标音频数据
S302

图5

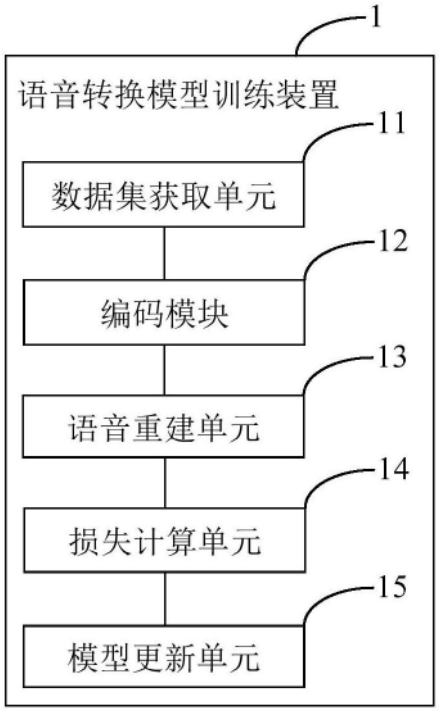

1
语音转换模型训练装置
11
数据集获取单元
12
编码模块
13
语音重建单元
14
损失计算单元
15
模型更新单元

图6

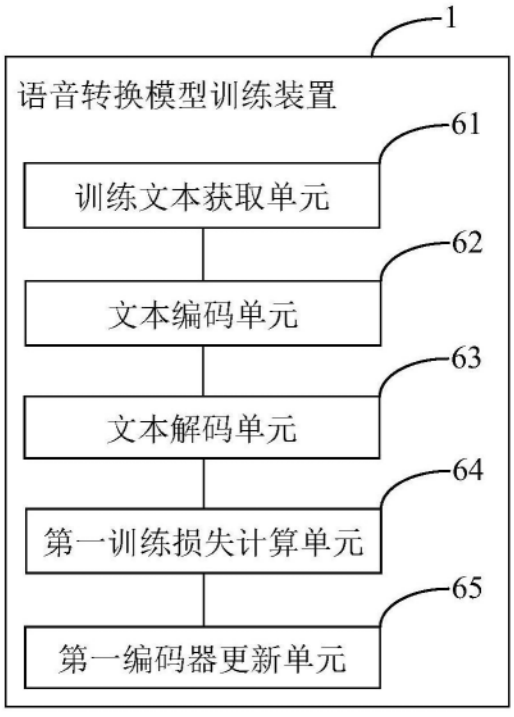

1
语音转换模型训练装置
61
训练文本获取单元
62
文本编码单元
63
文本解码单元
64
第一训练损失计算单元
65
第一编码器更新单元

图7

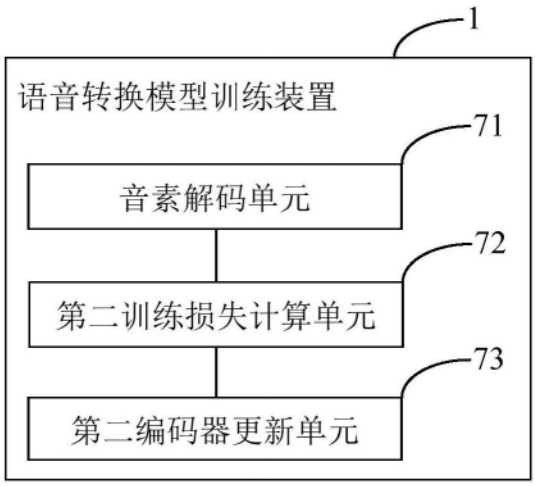

1
语音转换模型训练装置
71
音素解码单元
72
第二训练损失计算单元
73
第二编码器更新单元

图8

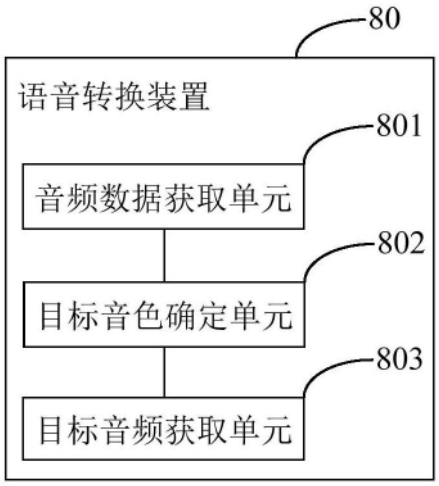

80
语音转换装置
801
音频数据获取单元
802
目标音色确定单元
803
目标音频获取单元

图9

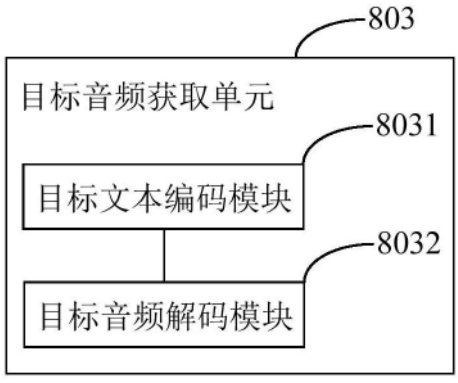

803
目标音频获取单元
8031
目标文本编码模块
8032
目标音频解码模块

图10

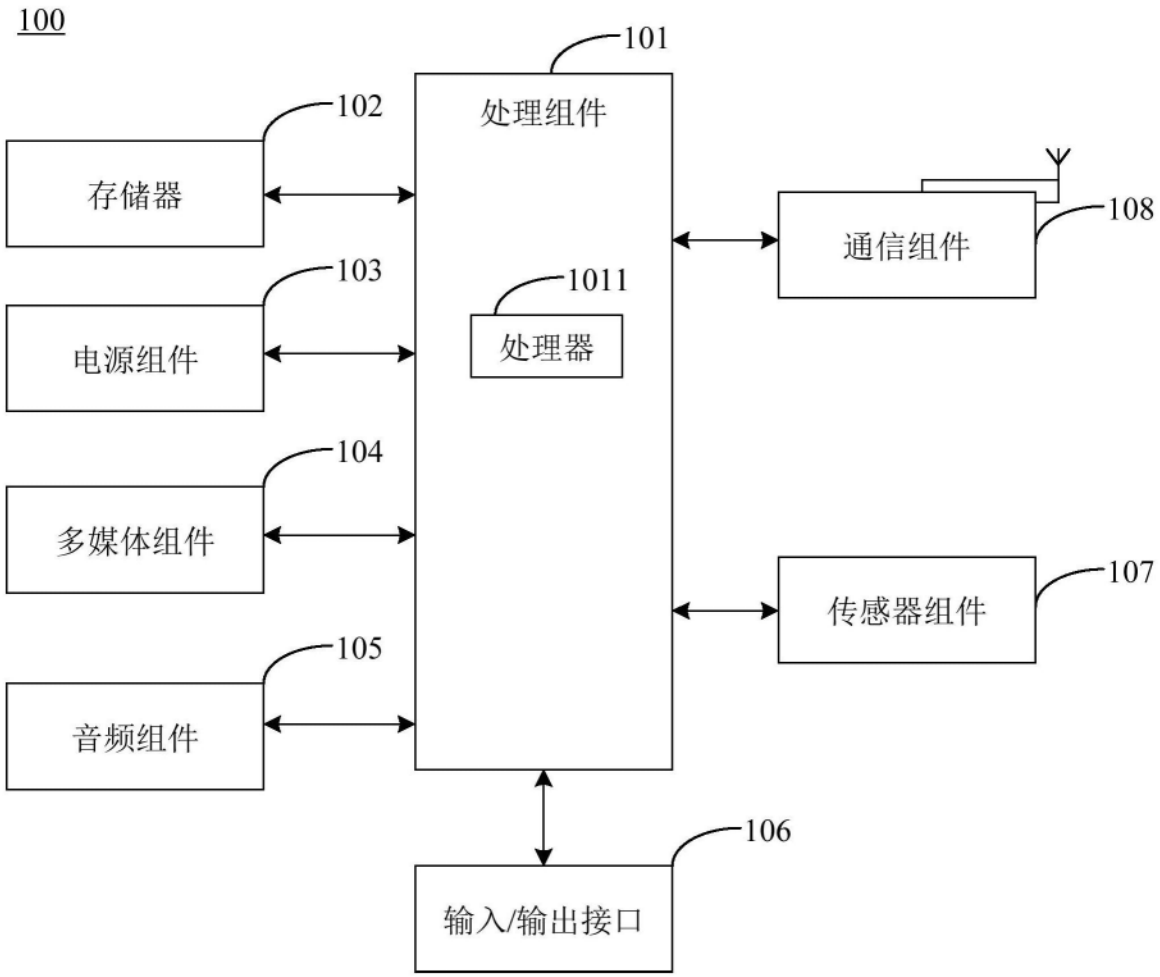

100
101
处理组件
102
存储器
103
电源组件
1011
处理器
104
多媒体组件
105
音频组件
108
通信组件
107
传感器组件
106
输入/输出接口

图11