(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114882862 A
(43)申请公布日 2022.08.09

(21)申请号 202210468926.8

(22)申请日 2022.04.29

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 邓利群 朱杰明 张立超 赵洲

(74)专利代理机构 深圳市深佳知识产权代理事务所(普通合伙) 44285
专利代理师 吴欣蔚

(51)Int.Cl.
*G10L 13/02*(2013.01)
*G10L 13/033*(2013.01)
*G10L 15/02*(2006.01)
*G10L 15/08*(2006.01)

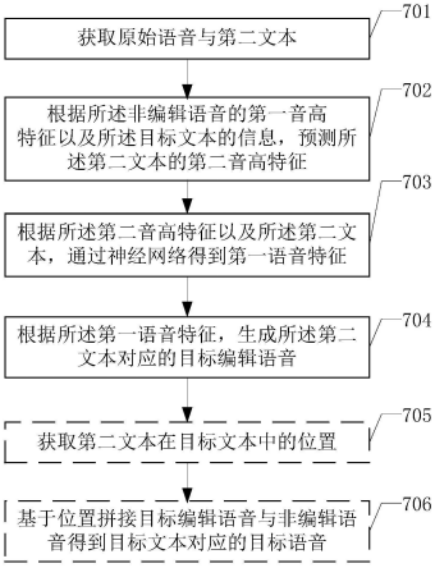
权利要求书3页 说明书31页 附图12页

(54)发明名称

一种语音处理方法及相关设备

(57)摘要

一种语音处理方法,应用于歌声编辑领域,所述方法包括:获取原始语音以及第二文本;根据原始语音中非编辑语音的第一音高特征以及目标文本的信息,预测所述第二文本的第二音高特征;根据所述第二音高特征以及所述第二文本,通过神经网络得到所述第二文本对应的第一语音特征;根据所述第一语音特征,生成所述第二文本对应的目标编辑语音。本申请通过预测第二文本(待编辑文本)的音高特征,根据音高特征生成第二文本的第一语音特征,并基于第一语音特征生成第二文本对应目标编辑语音,使得歌声编辑前后的语音的音高特征相似,进而实现目标编辑语音的听感与原始语音的听感目标编辑语音的听感与原始语音的听感类似。




CN 114882862 A

1.一种语音处理方法,其特征在于,所述方法包括:

获取原始语音以及第二文本,所述第二文本为目标文本中除了第一文本以外的文本,所述目标文本与所述原始语音对应的原始文本都包括所述第一文本,所述第一文本在所述原始语音中对应的语音为非编辑语音;

根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息,预测所述第二文本的第二音高特征;

根据所述第二音高特征以及所述第二文本,通过神经网络得到所述第二文本对应的第一语音特征;

根据所述第一语音特征,生成所述第二文本对应的目标编辑语音。

2.根据权利要求1所述的方法,其特征在于,所述原始语音的内容为用户的歌声。

3.根据权利要求1至3任一所述的方法,其特征在于,所述根据所述非编辑语音的第一音高(pitch)特征以及所述第二文本包括:

根据所述非编辑语音的第一音高(pitch)特征、所述目标文本的信息以及所述非编辑语音的第二语音特征;所述第二语音特征携带有如下信息的至少一种:

所述非编辑语音的部分语音帧或全部语音帧;

所述非编辑语音的声纹特征;

所述非编辑语音的音色特征;

所述非编辑语音的韵律特征;以及,

所述非编辑语音的节奏特征。

4.根据权利要求1至4任一所述的方法,其特征在于,所述目标文本的信息,包括:

所述目标文本中各个音素的文本嵌入(text embedding)。

5.根据权利要求1至5任一所述的方法,其特征在于,所述目标文本为将所述第二文本插入到所述第一文本得到的文本;或者,所述目标文本为将所述第一文本的第一部分文本删除得到的文本,所述第二文本为与所述第一部分文本相邻的文本;

所述根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息,预测所述第二文本的第二音高特征,包括:

将所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息进行融合,以得到第一融合结果;

将所述第一融合结果输入到第二神经网络,得到所述第二文本的第二音高特征。

6.根据权利要求1至5任一所述的方法,其特征在于,所述目标文本为将所述第一文本中的第二部分文本替换为所述第二文本得到的;

所述根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息,预测所述第二文本的第二音高特征,包括:

将所述非编辑语音的第一音高(pitch)特征输入到第三神经网络,得到初始音高特征,所述第一初始音高特征包括多个帧中每个帧的音高;

将所述目标文本的信息输入到第四神经网络,得到所述第二文本的发音特征,所述发音特征用于指示所述初始音高特征包括的多个帧中各个帧是否发音;

将所述初始音高特征和所述发音特征进行融合,以得到所述第二文本的第二音高特征。

7.根据权利要求1至6任一所述的方法,其特征在于,所述方法还包括:

根据所述非编辑语音中各个音素的帧数以及所述目标文本的信息,预测所述第二文本中各个音素的帧数。

8.根据权利要求1至7任一所述的方法,其特征在于,所述第一音高(pitch)特征,包括:所述非编辑语音的多帧中的每一帧的音高特征;

所述第二音高特征,包括:所述目标编辑语音的多帧中的每一帧的音高特征。

9.根据权利要求7或8所述的方法,其特征在于,所述根据所述非编辑语音中各个音素的帧数以及所述目标文本的信息,包括:

根据所述非编辑语音中各个音素的帧数、所述目标文本的信息以及所述非编辑语音的第二语音特征。

10.根据权利要求1至9任一所述的方法,其特征在于,所述方法还包括:

获取所述第二文本在所述目标文本中的位置;

基于所述位置拼接所述目标编辑语音与所述非编辑语音得到所述目标文本对应的目标语音。

11.一种语音处理装置,其特征在于,所述装置包括:

获取模块,用于获取原始语音以及第二文本,所述第二文本为目标文本中除了第一文本以外的文本,所述目标文本与所述原始语音对应的原始文本都包括所述第一文本,所述第一文本在所述原始语音中对应的语音为非编辑语音;

音高预测模块,用于根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息,预测所述第二文本的第二音高特征;

生成模块,用于根据所述第二音高特征以及所述第二文本,通过神经网络得到所述第二文本对应的第一语音特征;

根据所述第一语音特征,生成所述第二文本对应的目标编辑语音。

12.根据权利要求11所述的装置,其特征在于,所述原始语音的内容为用户的歌声。

13.根据权利要求11或12所述的装置,其特征在于,所述根据所述非编辑语音的第一音高(pitch)特征以及所述第二文本包括:

根据所述非编辑语音的第一音高(pitch)特征、所述目标文本的信息以及所述非编辑语音的第二语音特征;所述第二语音特征携带有如下信息的至少一种:

所述非编辑语音的部分语音帧或全部语音帧;

所述非编辑语音的声纹特征;

所述非编辑语音的音色特征;

所述非编辑语音的韵律特征;以及,

所述非编辑语音的节奏特征。

14.根据权利要求11至14任一所述的装置,其特征在于,所述目标文本的信息,包括:所述目标文本中各个音素的文本嵌入(text embedding)。

15.根据权利要求11至14任一所述的装置,其特征在于,所述目标文本为将所述第二文本插入到所述第一文本得到的文本;或者,所述目标文本为将所述第一文本的第一部分文本删除得到的文本,所述第二文本为与所述第一部分文本相邻的文本;

所述音高预测模块,具体用于:

将所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息进行融合,以得到第一融合结果;

将所述第一融合结果输入到第二神经网络,得到所述第二文本的第二音高特征。

16.根据权利要求11至14任一所述的装置,其特征在于,所述目标文本为将所述第一文本中的第二部分文本替换为所述第二文本得到的;

所述音高预测模块,具体用于:

将所述非编辑语音的第一音高(pitch)特征输入到第三神经网络,得到初始音高特征,所述第一初始音高特征包括多个帧中每个帧的音高;

将所述目标文本的信息输入到第四神经网络,得到所述第二文本的发音特征,所述发音特征用于指示所述初始音高特征包括的多个帧中各个帧是否发音;

将所述初始音高特征和所述发音特征进行融合,以得到所述第二文本的第二音高特征。

17.根据权利要求11至16任一所述的装置,其特征在于,所述装置还包括:

时长预测模块,用于根据所述非编辑语音中各个音素的帧数以及所述目标文本的信息,预测所述第二文本中各个音素的帧数。

18.根据权利要求11至17任一所述的装置,其特征在于,所述第一音高(pitch)特征,包括:所述非编辑语音的多帧中的每一帧的音高特征;

所述第二音高特征,包括:所述目标编辑语音的多帧中的每一帧的音高特征。

19.根据权利要求17或18所述的装置,其特征在于,所述时长预测模块,具体用于:

根据所述非编辑语音中各个音素的帧数、所述目标文本的信息以及所述非编辑语音的第二语音特征。

20.根据权利要求11至19任一所述的装置,其特征在于,所述获取模块还用于:

获取所述第二文本在所述目标文本中的位置;

所述生成模块,还用于基于所述位置拼接所述目标编辑语音与所述非编辑语音得到所述目标文本对应的目标语音。

21.一种语音处理设备,其特征在于,包括:处理器,所述处理器与存储器耦合,所述存储器用于存储程序或指令,当所述程序或指令被所述处理器执行时,使得所述语音处理设备执行如权利要求1至10中任一项所述的方法。

22.根据权利要求21所述的设备,其特征在于,所述设备还包括:

输入单元,用于接收第二文本;

输出单元,用于播放所述第二文本对应的目标编辑语音或者目标文本对应的目标语音。

23.一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指令,所述指令在计算机上执行时,使得所述计算机执行如权利要求1至10中任一项所述的方法。

24.一种计算机程序产品,其特征在于,所述计算机程序产品在计算机上执行时,使得所述计算机执行如权利要求1至10中任一项所述的方法。

# 一种语音处理方法及相关设备

## 技术领域

[0001] 本申请实施例涉及人工智能领域领域，尤其涉及一种语音处理方法及相关设备。

## 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、技术及应用系统。换句话说，人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式作出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。人工智能领域的研究包括机器人，自然语言处理，计算机视觉，决策与推理，人机交互，推荐与搜索，AI基础理论等。

[0003] 目前，语音编辑具有非常重要的实用意义。比如，在用户录制歌曲(例如清唱)等场景下，经常会由于口误而导致语音中的某些内容出错。该种情况下，语音编辑便可帮助用户快速地修正原始歌声中的错误内容，生成校正后的语音。常用的语音编辑方法是通过预先构建含有大量语音片段的数据库，从数据库中获取发音单元的片段，并用该片段替换原始语音中的错误片段，进而生成校正后的语音。

[0004] 然而，上述语音编辑的方式依赖数据库中语音片段的多样性，在数据库中语音片段较少的情况下，会导致校正后的语音(例如用户的歌声)的听感较差。

## 发明内容

[0005] 本申请实施例提供了一种语音处理方法及相关设备，可以实现编辑歌声的听感与原始语音的听感类似，提升用户体验。

[0006] 第一方面，本申请提供了一种语音处理方法，可以应用于用户录制短视频、老师录制授课语音等场景。该方法可以由语音处理设备执行，也可以由语音处理设备的部件(例如处理器、芯片、或芯片系统等)执行。其中，该语音处理设备可以是终端设备也可以是云端设备，所述方法包括：获取原始语音以及第二文本，所述第二文本为目标文本中除了第一文本以外的文本，所述目标文本与所述原始语音对应的原始文本都包括所述第一文本，所述第一文本在所述原始语音中对应的语音为非编辑语音；根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息，预测所述第二文本的第二音高特征；根据所述第二音高特征以及所述第二文本，通过神经网络得到所述第二文本对应的第一语音特征；根据所述第一语音特征，生成所述第二文本对应的目标编辑语音。本申请通过预测第二文本(待编辑文本)的音高特征，根据音高特征生成第二文本的第一语音特征，并基于第一语音特征生成第二文本对应目标编辑语音，使得歌声编辑前后的语音的音高特征相似，进而实现目标编辑语音的听感与原始语音的听感目标编辑语音的听感与原始语音的听感类似。

[0007] 另外，获取第二文本的方式有多种，可以是直接获取第二文本；也可以是先获取位置信息(也可以理解为是标记信息，用于指示第二文本在目标文本中的位置)，在根据位置

与目标文本获取第二文本，位置信息用于表示第二文本在目标文本中的位置；还可以是获取目标文本与原始文本(或者获取目标文本与原始语音，对原始语音进行识别得到原始文本)，再基于原始文本与目标文本确定第二文本。

[0008] 在一种可能的实现中，基于第二语音特征生成第二文本对应的目标编辑语音，包括：基于第二语音特征，通过声码器，生成目标编辑语音。

[0009] 该种可能的实现方式中，根据声码器将第二语音特征转化为目标编辑语音，进而使得目标编辑语音具有与原始语音相近的语音特征，提升用户的听感。

[0010] 在一种可能的实现中，所述原始语音的内容为用户的歌声，例如可以为用户清唱时录制的语音。

[0011] 在一种可能的实现中，获取原始语音与第二文本，包括：接收终端设备发送的原始语音与第二文本；方法还包括：向终端设备发送目标编辑语音，目标编辑语音用于终端设备生成目标文本对应的目标语音。也可以理解为是交互场景，由云端设备进行复杂的计算操作，由终端设备执行简单的拼接操作，从终端设备处获取原始语音与第二文本，云端设备生成目标编辑语音之后，向终端设备发送目标编辑语音，再由终端设备进行拼接得到目标语音。

[0012] 该种可能的实现方式中，在语音处理设备是云端设备的情况下，一方面，可以通过云端设备与终端设备的交互，由云端设备进行复杂的计算得到目标编辑语音并返给终端设备，可以减少终端设备的算力与存储空间。另一方面，可以根据原始语音中非编辑区域的语音特征生成修改文本对应的目标编辑语音，进而与非编辑语音生成目标文本对应的目标语音。

[0013] 可选地，在第一方面的一种可能的实现方式中，上述步骤：获取原始语音与第二文本，包括：接收终端设备发送的原始语音与目标文本；方法还包括：基于非编辑语音与目标编辑语音生成目标文本对应的目标语音，向终端设备发送目标语音。

[0014] 该种可能的实现方式中，接收终端设备发送的原始语音与目标文本，可以获取非编辑语音，并根据非编辑语音的第一语音特征生成第二文本对应的第二语音特征，进而根据声码器得到目标编辑语音，并拼接目标编辑语音与非编辑语音生成目标语音。相当于，处理过程都在语音处理设备，结果返回给终端设备。由云端设备进行复杂的计算得到目标语音并返给终端设备，可以减少终端设备的算力与存储空间。

[0015] 在一种可能的实现中，所述根据所述非编辑语音的第一音高(pitch)特征以及所述第二文本包括：根据所述非编辑语音的第一音高(pitch)特征、所述目标文本的信息以及所述非编辑语音的第二语音特征；所述第二语音特征携带有如下信息的至少一种：所述非编辑语音的部分语音帧或全部语音帧；所述非编辑语音的声纹特征；所述非编辑语音的音色特征；所述非编辑语音的韵律特征；以及，所述非编辑语音的节奏特征。

[0016] 其中，第一语音特征可以与第二语音特征的韵律、音色和/或信噪比等相同或相近，韵律可以反映出发音者的情感状态或讲话形式等，韵律泛指语调、音调、重音强调、停顿或节奏等特征。

[0017] 在一种可能的实现中，第二语音特征携带有原始语音的声纹特征。其中，获取声纹特征的方式可以是直接获取，也可以是通过识别原始语音得到该声纹特征等。

[0018] 该种可能的实现方式中，一方面，通过引入原始语音的声纹特征，使得后续生成的

第一语音特征也携带有该原始语音的声纹特征，进而提升目标编辑语音与原始语音的相近程度。另一方面，在发音者(或者用户)的数量为多个的情况下，引入声纹特征可以提升后续预测的语音特征更加与原始语音的发音者的声纹相似。

[0019] 在一种可能的实现中，所述目标文本的信息，包括：

[0020] 所述目标文本中各个音素的文本嵌入(text embedding)。

[0021] 在一种可能的实现中，所述目标文本为将所述第二文本插入到所述第一文本得到的文本；或者，所述目标文本为将所述第一文本的第一部分文本删除得到的文本，所述第二文本为与所述第一部分文本相邻的文本；

[0022] 所述根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息，预测所述第二文本的第二音高特征，包括：

[0023] 将所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息进行融合，以得到第一融合结果；

[0024] 将所述第一融合结果输入到第二神经网络，得到所述第二文本的第二音高特征。

[0025] 在一种可能的实现中，所述目标文本为将所述第一文本中的第二部分文本替换为所述第二文本得到的；

[0026] 所述根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息，预测所述第二文本的第二音高特征，包括：

[0027] 将所述非编辑语音的第一音高(pitch)特征输入到第三神经网络，得到初始音高特征，所述第一初始音高特征包括多个帧中每个帧的音高；

[0028] 将所述目标文本的信息输入到第四神经网络，得到所述第二文本的发音特征，所述发音特征用于指示所述初始音高特征包括的多个帧中各个帧是否发音；

[0029] 将所述初始音高特征和所述发音特征进行融合，以得到所述第二文本的第二音高特征。

[0030] 在一种可能的实现中，所述方法还包括：

[0031] 根据所述非编辑语音中各个音素的帧数以及所述目标文本的信息，预测所述第二文本中各个音素的帧数。

[0032] 在一种可能的实现中，所述第一音高(pitch)特征，包括：所述非编辑语音的多帧中的每一帧的音高特征；

[0033] 所述第二音高特征，包括：所述目标编辑语音的多帧中的每一帧的音高特征。

[0034] 在一种可能的实现中，所述根据所述非编辑语音中各个音素的帧数以及所述目标文本的信息，包括：

[0035] 根据所述非编辑语音中各个音素的帧数、所述目标文本的信息以及所述非编辑语音的第二语音特征。

[0036] 在一种可能的实现中，上述步骤还包括：获取第二文本在目标文本中的位置；基于位置拼接目标编辑语音与非编辑语音得到目标文本对应的目标语音。也可以理解为是用目标编辑语音替换原始语音中的编辑语音，该编辑语音为原始语音中除了非编辑语音以外的语音。

[0037] 该种可能的实现方式中，可以根据第二文本在目标文本中的位置拼接目标编辑语音与非编辑语音。如果第一文本是原始文本与目标文本中的所有重叠文本，则可以在不改

变原始语音中非编辑语音的情况下生成所需文本(即目标文本)的语音。

[0038] 可选地,在第一方面的一种可能的实现方式中,上述步骤还包括:基于目标文本、原始文本以及原始语音确定非编辑语音,具体可以是:基于目标文本与原始文本确定第一文本;基于第一文本、原始文本与原始语音确定非编辑语音。

[0039] 该种可能的实现方式中,通过对比原始文本与原始语音,确定第一文本在原始语音中的非编辑语音,便于后续第一语音特征的生成。

[0040] 可选地,在第一方面的一种可能的实现方式中,上述步骤:基于目标文本与原始文本确定第一文本,包括:基于目标文本与原始文本确定重叠文本;向用户显示重叠文本;响应用户的第二操作,从重叠文本中确定第一文本。

[0041] 第二方面,本申请提供了一种语音处理装置,所述装置包括:

[0042] 获取模块,用于获取原始语音以及第二文本,所述第二文本为目标文本中除了第一文本以外的文本,所述目标文本与所述原始语音对应的原始文本都包括所述第一文本,所述第一文本在所述原始语音中对应的语音为非编辑语音;

[0043] 音高预测模块,用于根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息,预测所述第二文本的第二音高特征;

[0044] 生成模块,用于根据所述第二音高特征以及所述第二文本,通过神经网络得到所述第二文本对应的第一语音特征;

[0045] 根据所述第一语音特征,生成所述第二文本对应的目标编辑语音。

[0046] 在一种可能的实现中,所述原始语音的内容为用户的歌声。

[0047] 在一种可能的实现中,所述根据所述非编辑语音的第一音高(pitch)特征以及所述第二文本包括:

[0048] 根据所述非编辑语音的第一音高(pitch)特征、所述目标文本的信息以及所述非编辑语音的第二语音特征;所述第二语音特征携带有如下信息的至少一种:

[0049] 所述非编辑语音的部分语音帧或全部语音帧;

[0050] 所述非编辑语音的声纹特征;

[0051] 所述非编辑语音的音色特征;

[0052] 所述非编辑语音的韵律特征;以及,

[0053] 所述非编辑语音的节奏特征。

[0054] 在一种可能的实现中,所述目标文本的信息,包括:所述目标文本中各个音素的文本嵌入(text embedding)。

[0055] 在一种可能的实现中,所述目标文本为将所述第二文本插入到所述第一文本得到的文本;或者,所述目标文本为将所述第一文本的第一部分文本删除得到的文本,所述第二文本为与所述第一部分文本相邻的文本;

[0056] 所述音高预测模块,具体用于:

[0057] 将所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息进行融合,以得到第一融合结果;

[0058] 将所述第一融合结果输入到第二神经网络,得到所述第二文本的第二音高特征。

[0059] 在一种可能的实现中,所述目标文本为将所述第一文本中的第二部分文本替换为所述第二文本得到的;

[0060] 所述音高预测模块,具体用于:

[0061] 将所述非编辑语音的第一音高(pitch)特征输入到第三神经网络,得到初始音高特征,所述第一初始音高特征包括多个帧中每个帧的音高;

[0062] 将所述目标文本的信息输入到第四神经网络,得到所述第二文本的发音特征,所述发音特征用于指示所述初始音高特征包括的多个帧中各个帧是否发音;

[0063] 将所述初始音高特征和所述发音特征进行融合,以得到所述第二文本的第二音高特征。

[0064] 在一种可能的实现中,所述装置还包括:

[0065] 时长预测模块,用于根据所述非编辑语音中各个音素的帧数以及所述目标文本的信息,预测所述第二文本中各个音素的帧数。

[0066] 在一种可能的实现中,所述第一音高(pitch)特征,包括:所述非编辑语音的多帧中的每一帧的音高特征;

[0067] 所述第二音高特征,包括:所述目标编辑语音的多帧中的每一帧的音高特征。

[0068] 在一种可能的实现中,所述时长预测模块,具体用于:

[0069] 根据所述非编辑语音中各个音素的帧数、所述目标文本的信息以及所述非编辑语音的第二语音特征。

[0070] 在一种可能的实现中,所述获取模块还用于:

[0071] 获取所述第二文本在所述目标文本中的位置;

[0072] 所述生成模块,还用于基于所述位置拼接所述目标编辑语音与所述非编辑语音得到所述目标文本对应的目标语音。

[0073] 本申请第三方面提供了一种语音处理设备,该语音处理设备执行前述第一方面或第一方面的任意可能的实现方式中的方法。

[0074] 本申请第四方面提供了一种语音处理设备,包括:处理器,处理器与存储器耦合,存储器用于存储程序或指令,当程序或指令被处理器执行时,使得该语音处理设备实现上述第一方面或第一方面的任意可能的实现方式中的方法。

[0075] 本申请第五方面提供了一种计算机可读介质,其上存储有计算机程序或指令,当计算机程序或指令在计算机上运行时,使得计算机执行前述第一方面或第一方面的任意可能的实现方式中的方法。

[0076] 本申请第六方面提供了一种计算机程序产品,该计算机程序产品在计算机上执行时,使得计算机执行前述第一方面或第一方面的任意可能的实现方式中的方法。

## 附图说明

[0077] 图1为本申请提供的一种系统架构的结构示意图;

[0078] 图2为本申请提供的一种卷积神经网络结构示意图;

[0079] 图3为本申请提供的另一种卷积神经网络结构示意图;

[0080] 图4为本申请提供的一种芯片硬件结构示意图;

[0081] 图5为本申请提供的一种神经网络的训练方法的示意性流程图;

[0082] 图6为本申请提供的一种神经网络的结构示意图;

[0083] 图7a为本申请提供的语音处理方法一个流程示意图;

[0084] 图7b为本申请提供的一个时长预测示意图；

[0085] 图7c为本申请提供的一个音高预测示意图；

[0086] 图7d为本申请提供的一个音高预测示意图；

[0087] 图8-图10为本申请提供的语音处理设备显示界面的几种示意图；

[0088] 图11为本申请提供的一种双向解码器的结构示意图；

[0089] 图12为本申请提供的语音处理设备显示界面的另一种示意图；

[0090] 图13为本申请提供的语音处理方法另一个流程示意图；

[0091] 图14-图16本申请提供的语音处理设备的几种结构示意图。

## 具体实施方式

[0092] 本申请实施例提供了一种语音处理方法及相关设备，可以实现编辑语音的听感与原始语音的听感类似，提升用户体验。

[0093] 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描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获取的所有其他实施例，都属于本申请保护的范围。

[0094] 为了便于理解，下面先对本申请实施例主要涉及的相关术语和概念进行介绍。

[0095] 1、神经网络

[0096] 神经网络可以是由神经单元组成的，神经单元可以是指以$X_s$和截距1为输入的运算单元，该运算单元的输出可以为：

[0097] $$h_{W,b}(x)=f\left(W^{T}x\right)=f(\sum_{s=1}^{n}W_{s}x_{s}+b)$$。

[0098] 其中，s＝1、2、……n，n为大于1的自然数，$W_s$为$X_s$的权重，b为神经单元的偏置。f为神经单元的激活函数(activation functions)，用于将非线性特性引入神经网络中，来将神经单元中的输入信号转换为输出信号。该激活函数的输出信号可以作为下一层卷积层的输入。激活函数可以是sigmoid函数。神经网络是将许多个上述单一的神经单元联结在一起形成的网络，即一个神经单元的输出可以是另一个神经单元的输入。每个神经单元的输入可以与前一层的局部接受域相连，来提取局部接受域的特征，局部接受域可以是由若干个神经单元组成的区域。

[0099] 2、深度神经网络

[0100] 深度神经网络(deep neural network，DNN)，也称多层神经网络，可以理解为具有很多层隐含层的神经网络，这里的“很多”并没有特别的度量标准。从DNN按不同层的位置划分，DNN内部的神经网络可以分为三类：输入层，隐含层，输出层。一般来说第一层是输入层，最后一层是输出层，中间的层数都是隐含层。层与层之间是全连接的，也就是说，第i层的任意一个神经元一定与第i+1层的任意一个神经元相连。当然，深度神经网络也可能不包括隐藏层，具体此处不做限定。

[0101] 深度神经网络中的每一层的工作可以用数学表达式$\vec{y}=\alpha\left(W\vec{x}+\vec{b}\right)$来描述：从物理层面深度神经网络中的每一层的工作可以理解为通过五种对输入空间(输入向量的集合)的操作，完成输入空间到输出空间的变换(即矩阵的行空间到列空间)，这五种操作包

括:1、升维/降维;2、放大/缩小;3、旋转;4、平移;5、“弯曲”。其中1、2、3的操作由$Wx$完成,4的操作由$+\vec{b}$完成,5的操作则由α()来实现。这里之所以用“空间”二字来表述是因为被分类的对象并不是单个事物,而是一类事物,空间是指这类事物所有个体的集合。其中,W是权重向量,该向量中的每一个值表示该层神经网络中的一个神经元的权重值。该向量W决定着上文所述的输入空间到输出空间的空间变换,即每一层的权重W控制着如何变换空间。训练深度神经网络的目的,也就是最终获取训练好的神经网络的所有层的权重矩阵(由很多层的向量W形成的权重矩阵)。因此,神经网络的训练过程本质上就是学习控制空间变换的方式,更具体的就是学习权重矩阵。

[0102] 3、卷积神经网络

[0103] 卷积神经网络(convolutional neuron network,CNN)是一种带有卷积结构的深度神经网络。卷积神经网络包含了一个由卷积层和子采样层构成的特征抽取器。该特征抽取器可以看作是滤波器,卷积过程可以看作是使同一个可训练的滤波器与一个输入的图像或者卷积特征平面(feature map)做卷积。卷积层是指卷积神经网络中对输入信号进行卷积处理的神经元层。在卷积神经网络的卷积层中,一个神经元可以只与部分邻层神经元连接。一个卷积层中,通常包含若干个特征平面,每个特征平面可以由一些矩形排列的神经单元组成。同一特征平面的神经单元共享权重,这里共享的权重就是卷积核。共享权重可以理解为提取图像信息的方式与位置无关。这其中隐含的原理是:图像的某一部分的统计信息与其他部分是一样的。即意味着在某一部分学习的图像信息也能用在另一部分上。所以对于图像上的所有位置,都能使用同样的学习获取的图像信息。在同一卷积层中,可以使用多个卷积核来提取不同的图像信息,一般地,卷积核数量越多,卷积操作反映的图像信息越丰富。

[0104] 卷积核可以以随机大小的矩阵的形式初始化,在卷积神经网络的训练过程中卷积核可以通过学习获取合理的权重。另外,共享权重带来的直接好处是减少卷积神经网络各层之间的连接,同时又降低了过拟合的风险。本申请实施例中的分离网络、识别网络、检测网络、深度估计网络等网络都可以是CNN。

[0105] 4、循环神经网络(RNN)

[0106] 在传统的神经网络中模型中,层与层之间是全连接的,每层之间的节点是无连接的。但是这种普通的神经网络对于很多问题是无法解决的。比如,预测句子的下一个单词是什么,因为一个句子中前后单词并不是独立的,一般需要用到前面的单词。循环神经网络(recurrent neural network,RNN)指的是一个序列当前的输出与之前的输出也有关。具体的表现形式为网络会对前面的信息进行记忆,保存在网络的内部状态中,并应用于当前输出的计算中。

[0107] 5、损失函数

[0108] 在训练深度神经网络的过程中,因为希望深度神经网络的输出尽可能的接近真正想要预测的值,所以可以通过比较当前网络的预测值和真正想要的目标值,再根据两者之间的差异情况来更新每一层神经网络的权重向量(当然,在第一次更新之前通常会有初始化的过程,即为深度神经网络中的各层预先配置参数),比如,如果网络的预测值高了,就调整权重向量让它预测低一些,不断的调整,直到神经网络能够预测出真正想要的目标值。因此,就需要预先定义“如何比较预测值和目标值之间的差异”,这便是损失函数(loss

function)或目标函数(objective function),它们是用于衡量预测值和目标值的差异的重要方程。其中,以损失函数举例,损失函数的输出值(loss)越高表示差异越大,那么深度神经网络的训练就变成了尽可能缩小这个loss的过程。

[0109] 6、从文本到语音

[0110] 从文本到语音(text to speech,TTS)是将文本转换成语音的程序或软件系统。

[0111] 7、声码器

[0112] 声码器是一种声音信号处理模块或软件,可以将声学特征编码生成声音波形。

[0113] 8、音高

[0114] 音高也可以称之为基频,当发声体由于振动而发出声音时,声音一般可以分解为许多单纯的正弦波,也就是说所有的自然声音基本都是由许多频率不同的正弦波组成的,其中频率最低的正弦波即为基音(即基频,可以用F0表示),而其他频率较高的正弦波则为泛音。

[0115] 9、韵律

[0116] 语音合成领域中,韵律泛指控制语调、音调、重音强调、停顿和节奏等功能的特征。韵律可以反映出说话者的情感状态或讲话形式等。

[0117] 10、音素

[0118] 音素(phone):是根据语音的自然属性划分出来的最小语音单位,依据音节里的发音动作来分析,一个动作构成一个音素。音素分为元音与辅音两大类。例如,汉语音节a(例如,一声:啊)只有一个音素,ai(例如四声:爱)有两个音素,dai(例如一声:呆)有三个音素等。

[0119] 11、词向量(embedding)

[0120] 词向量也可以称为“词嵌入”、“向量化”、“向量映射”、“嵌入”等。从形式上讲,词向量是用一个稠密的向量表示一个对象。

[0121] 12、语音特征

[0122] 语音特征:将经过处理的语音信号转换成一种简洁而有逻辑的表示形式,比实际信号更有鉴别性和可靠性。在获取一段语音信号后,可以从语音信号中提取语音特征。其中,提取方法通常为每个语音信号提取一个多维特征向量。语音信号的参数化表示方法有很多种,例如:感知线性预测(perceptual linear predictive,PLP)、线性预测编码(linear predictive coding,LPC)和频率倒谱系数(mel frequency cepstrum coefficient,MFCC)等。

[0123] 13、transformer层

[0124] 神经网络包括嵌入层和至少一个transformer层,至少一个transformer层可以为N个transformer层(N大于0的整数),其中,每个transformer层包括依次相邻的注意力层、加和与归一化(add&norm)层、前馈(feed forward)层和加和与归一化层。在嵌入层,对当前输入进行嵌入处理,得到多个特征向量;在所述注意力层,从所述第一transformer层的上一层获取P个输入向量,以P个输入向量中的任意的第一输入向量为中心,基于预设的注意力窗口范围内的各个输入向量与该第一输入向量之间的关联度,得到该第一输入向量对应的中间向量,如此确定出P个输入向量对应的P个中间向量;在所述池化层,将所述P个中间向量合并为Q个输出向量,其中transformer层中最后一个transformer层得到的多个输出

向量用作所述当前输入的特征表示。

[0125] 接下来,结合具体例子对上述各步骤进行具体介绍。

[0126] 首先,在所述嵌入层,对当前输入进行嵌入处理,得到多个特征向量。

[0127] 嵌入层可以称为输入嵌入(input embedding)层。当前输入可以为文本输入,例如可以为一段文本,也可以为一个句子。文本可以为中文文本,也可以为英文文本,还可以为其他语言文本。嵌入层在获取当前输入后,可以对该当前输入中各个词进行嵌入处理,可得到各个词的特征向量。在一些实施例中,所述嵌入层包括输入嵌入层和位置编码(positional encoding)层。在输入嵌入层,可以对当前输入中的各个词进行词嵌入处理,从而得到各个词的词嵌入向量。在位置编码层,可以获取各个词在该当前输入中的位置,进而对各个词的位置生成位置向量。在一些示例中,各个词的位置可以为各个词在该当前输入中的绝对位置。以当前输入为“几号应还花呗”为例,其中的“几”的位置可以表示为第一位,“号”的位置可以表示为第二位,……。在一些示例中,各个词的位置可以为各个词之间的相对位置。仍以当前输入为“几号应还花呗”为例,其中的“几”的位置可以表示为“号”之前,“号”的位置可以表示为“几”之后、“应”之前,……。当得到当前输入中各个词的词嵌入向量和位置向量时,可以将各个词的位置向量和对应的词嵌入向量进行组合,得到各个词特征向量,即得到该当前输入对应的多个特征向量。多个特征向量可以表示为具有预设维度的嵌入矩阵。可以设定该多个特征向量中的特征向量个数为M,预设维度为H维,则该多个特征向量可以表示为M×H的嵌入矩阵。

[0128] 14、注意力机制(attention mechanism)

[0129] 注意力机制模仿了生物观察行为的内部过程,即一种将内部经验和外部感觉对齐从而增加部分区域的观察精细度的机制,能够利用有限的注意力资源从大量信息中快速筛选出高价值信息。注意力机制可以快速提取稀疏数据的重要特征,因而被广泛用于自然语言处理任务,特别是机器翻译。而自注意力机制(self-attention mechanism)是注意力机制的改进,其减少了对外部信息的依赖,更擅长捕捉数据或特征的内部相关性。注意力机制的本质思想可以改写为如下公式:

[0130] 其中,Lx=||Source||代表Source的长度,公式含义即将Source中的构成元素想象成是由一系列的数据对构成,此时给定目标Target中的某个元素Query,通过计算Query和各个Key的相似性或者相关性,得到每个Key对应Value的权重系数,然后对Value进行加权求和,即得到了最终的Attention数值。所以本质上Attention机制是对Source中元素的Value值进行加权求和,而Query和Key用来计算对应Value的权重系数。从概念上理解,把Attention可以理解为从大量信息中有选择地筛选出少量重要信息并聚焦到这些重要信息上,忽略大多不重要的信息。聚焦的过程体现在权重系数的计算上,权重越大越聚焦于其对应的Value值上,即权重代表了信息的重要性,而Value是其对应的信息。自注意力机制可以理解为内部Attention(intra attention),Attention机制发生在Target的元素Query和Source中的所有元素之间,自注意力机制指的是在Source内部元素之间或者Target内部元素之间发生的Attention机制,也可以理解为Target=Source这种特殊情况下的注意力计算机制,其具体计算过程是一样的,只是计算对象发生了变化而已。

[0131] 目前,语音编辑的场景越来越多,例如,歌声编辑的场景为用户在录制歌曲(例如清唱)等场景,为了修复由于口误带来的原始语音中的错误内容,通常会用到语音编辑。目

前的语音编辑方式是从数据库中获取语音片段，并用该语音片段替换错误内容，进而生成校正后的语音。

[0132] 然而，该种方式过于依赖数据库中存储的语音片段，若该语音片段与原始语音的音色、韵律、信噪比等相差较大，会导致校正后的人声前后不连贯、韵律不自然，导致校正后的语音听感较差。且虽然歌声编辑同语音编辑的场景非常相似，但不同于说话声的平稳语音，歌声数据在发音时长、声音能量和音高等维度上变化更大，现有的语音编辑技术难以直接应用于歌声编辑。

[0133] 为了解决上述问题，本申请提供一种语音编辑方法，在歌声编辑时，音高特征会影响目标编辑语音的听感与原始语音的听感，本申请通过预测第二文本(待编辑文本)的音高特征，根据音高特征生成第二文本的第一语音特征，并基于第一语音特征生成第二文本对应目标编辑语音，使得歌声编辑前后的语音的音高特征相似，进而实现目标编辑语音的听感与原始语音的听感目标编辑语音的听感与原始语音的听感类似。

[0134] 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描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获取的所有其他实施例，都属于本申请保护的范围。

[0135] 首先介绍本申请实施例提供的系统架构。

[0136] 参见附图1，本申请实施例提供了一种系统架构10。如所述系统架构10所示，数据采集设备16用于采集训练数据，本申请实施例中训练数据包括训练语音以及与该训练语音对应的训练文本。并将训练数据存入数据库13，训练设备12基于数据库13中维护的训练数据训练得到目标模型/规则101。下面将更详细地描述训练设备12如何基于训练数据得到目标模型/规则101，该目标模型/规则101能够用于实现本申请实施例提供的语音处理方法，即，将文本通过相关预处理后输入该目标模型/规则101，即可得到该文本的语音特征。本申请实施例中的目标模型/规则101具体可以为神经网络。需要说明的是，在实际的应用中，所述数据库13中维护的训练数据不一定都来自于数据采集设备16的采集，也有可能是从其他设备接收得到的。另外需要说明的是，训练设备12也不一定完全基于数据库13维护的训练数据进行目标模型/规则101的训练，也有可能从云端或其他地方获取训练数据进行模型训练，上述描述不应该作为对本申请实施例的限定。

[0137] 根据训练设备12训练得到的目标模型/规则101可以应用于不同的系统或设备中，如应用于图1所示的执行设备11，所述执行设备11可以是终端，如手机终端，平板电脑，笔记本电脑，AR/VR，车载终端等，还可以是服务器或者云端等。在附图1中，执行设备11配置有I/O接口112，用于与外部设备进行数据交互，用户可以通过客户设备14向I/O接口112输入数据，所述输入数据在本申请实施例中可以包括：第二语音特征、目标文本以及标记信息，输入数据也可以包括第二语音特征与第二文本。另外，输入数据可以是用户输入的，也可以是用户通过其他设备上传的，当然还可以来自数据库，具体此处不做限定。

[0138] 若输入数据包括第二语音特征、目标文本以及标记信息，则预处理模块113用于根据I/O接口112接收到的目标文本与标记信息进行预处理，在本申请实施例中，预处理模块113可以用于基于目标文本与标记信息确定目标文本中的目标编辑文本。若输入数据包括第二语音特征、第二文本，则预处理模块113用于根据I/O接口112接收到的目标文本与标记

信息进行预处理，例如，将目标文本转化为音素等准备工作。

[0139] 在执行设备11对输入数据进行预处理，或者在执行设备11的计算模块111执行计算等相关的处理过程中，执行设备11可以调用数据存储系统15中的数据、代码等以用于相应的处理，也可以将相应处理得到的数据、指令等存入数据存储系统15中。

[0140] 最后，I/O接口112将处理结果，如上述得到的第一语音特征返回给客户设备14，从而提供给用户。

[0141] 值得说明的是，训练设备12可以针对不同的目标或称不同的任务，基于不同的训练数据生成相应的目标模型/规则101，该相应的目标模型/规则101即可以用于实现上述目标或完成上述任务，从而为用户提供所需的结果或为后续的其他处理提供输入。

[0142] 在附图1中所示情况下，用户可以手动给定输入数据，该手动给定可以通过I/O接口112提供的界面进行操作。另一种情况下，客户设备14可以自动地向I/O接口112发送输入数据，如果要求客户设备14自动发送输入数据需要获得用户的授权，则用户可以在客户设备14中设置相应权限。用户可以在客户设备14查看执行设备11输出的结果，具体的呈现形式可以是显示、声音、动作等具体方式。客户设备14也可以作为数据采集端，采集如图所示输入I/O接口112的输入数据及输出I/O接口112的输出结果作为新的样本数据，并存入数据库13。当然，也可以不经过客户设备14进行采集，而是由I/O接口112直接将如图所示输入I/O接口112的输入数据及输出I/O接口112的输出结果，作为新的样本数据存入数据库13。

[0143] 值得注意的是，附图1仅是本申请实施例提供的一种系统架构的示意图，图中所示设备、器件、模块等之间的位置关系不构成任何限制，例如，在附图1中，数据存储系统15相对执行设备11是外部存储器，在其它情况下，也可以将数据存储系统15置于执行设备11中。

[0144] 如图1所示，根据训练设备12训练得到目标模型/规则101，该目标模型/规则101在本申请实施例中可以是神经网络，具体的，在本申请实施例提供的网络中，神经网络可以是循环神经网络、长短期记忆网络等。预测网络可以是卷积神经网络、循环神经网络等。

[0145] 可选地，本申请实施例中的神经网络与预测网络可以是单独的两个网络，也可以是一个多任务的神经网络，其中一个任务是输出时长，一个任务是预测音高特征，另外一个任务是输出语音特征。

[0146] 由于CNN是一种非常常见的神经网络，下面结合图2重点对CNN的结构进行详细的介绍。如前文的基础概念介绍所述，卷积神经网络是一种带有卷积结构的深度神经网络，是一种深度学习(deep learning)架构，深度学习架构是指通过机器学习的算法，在不同的抽象层级上进行多个层次的学习。作为一种深度学习架构，CNN是一种前馈(feed-forward)人工神经网络，该前馈人工神经网络中的各个神经元可以对输入其中的图像作出响应。

[0147] 如图2所示，卷积神经网络(CNN)100可以包括输入层110，卷积层/池化层120，以及神经网络层130其中池化层为可选的。

[0148] 卷积层/池化层120：

[0149] 卷积层：

[0150] 如图2所示卷积层/池化层120可以包括如示例121-126层，在一种实现中，121层为卷积层，122层为池化层，123层为卷积层，124层为池化层，125为卷积层，126为池化层；在另一种实现方式中，121、122为卷积层，123为池化层，124、125为卷积层，126为池化层。即卷积层的输出可以作为随后的池化层的输入，也可以作为另一个卷积层的输入以继续进行卷积

操作。

[0151] 以卷积层121为例,卷积层121可以包括很多个卷积算子,卷积算子也称为核,其在图像处理中的作用相当于一个从输入图像矩阵中提取特定信息的过滤器,卷积算子本质上可以是一个权重矩阵,这个权重矩阵通常被预先定义,在对图像进行卷积操作的过程中,权重矩阵通常在输入图像上沿着水平方向一个像素接着一个像素(或两个像素接着两个像素……这取决于步长stride的取值)的进行处理,从而完成从图像中提取特定特征的工作。该权重矩阵的大小应该与图像的大小相关,需要注意的是,权重矩阵的纵深维度(depth dimension)和输入图像的纵深维度是相同的,在进行卷积运算的过程中,权重矩阵会延伸到输入图像的整个深度。因此,和一个单一的权重矩阵进行卷积会产生一个单一纵深维度的卷积化输出,但是大多数情况下不使用单一权重矩阵,而是应用维度相同的多个权重矩阵。每个权重矩阵的输出被堆叠起来形成卷积图像的纵深维度。不同的权重矩阵可以用来提取图像中不同的特征,例如一个权重矩阵用来提取图像边缘信息,另一个权重矩阵用来提取图像的特定颜色,又一个权重矩阵用来对图像中不需要的噪点进行模糊化……该多个权重矩阵维度相同,经过该多个维度相同的权重矩阵提取后的特征图维度也相同,再将提取到的多个维度相同的特征图合并形成卷积运算的输出。

[0152] 这些权重矩阵中的权重值在实际应用中需要经过大量的训练得到,通过训练得到的权重值形成的各个权重矩阵可以从输入图像中提取信息,从而帮助卷积神经网络100进行正确的预测。

[0153] 当卷积神经网络100有多个卷积层的时候,初始的卷积层(例如121)往往提取较多的一般特征,该一般特征也可以称之为低级别的特征;随着卷积神经网络100深度的加深,越往后的卷积层(例如126)提取到的特征越来越复杂,比如高级别的语义之类的特征,语义越高的特征越适用于待解决的问题。

[0154] 池化层:

[0155] 由于常常需要减少训练参数的数量,因此卷积层之后常常需要周期性的引入池化层,即如图2中120所示例的121-126各层,可以是一层卷积层后面跟一层池化层,也可以是多层卷积层后面接一层或多层池化层。在图像处理过程中,池化层的唯一目的就是减少图像的空间大小。池化层可以包括平均池化算子和/或最大池化算子,以用于对输入图像进行采样得到较小尺寸的图像。平均池化算子可以在特定范围内对图像中的像素值进行计算产生平均值。最大池化算子可以在特定范围内取该范围内值最大的像素作为最大池化的结果。另外,就像卷积层中用权重矩阵的大小应该与图像大小相关一样,池化层中的运算符也应该与图像的大小相关。通过池化层处理后输出的图像尺寸可以小于输入池化层的图像的尺寸,池化层输出的图像中每个像素点表示输入池化层的图像的对应子区域的平均值或最大值。

[0156] 神经网络层130:

[0157] 在经过卷积层/池化层120的处理后,卷积神经网络100还不足以输出所需要的输出信息。因为如前所述,卷积层/池化层120只会提取特征,并减少输入图像带来的参数。然而为了生成最终的输出信息(所需要的类信息或别的相关信息),卷积神经网络100需要利用神经网络层130来生成一个或者一组所需要的类的数量的输出。因此,在神经网络层130中可以包括多层隐含层(如图2所示的131、132至13n)以及输出层140,该多层隐含层中所包

含的参数可以根据具体的任务类型的相关训练数据进行预先训练得到，例如该任务类型可以包括图像识别，图像分类，图像超分辨率重建等等。

[0158] 在神经网络层130中的多层隐含层之后，也就是整个卷积神经网络100的最后层为输出层140，该输出层140具有类似分类交叉熵的损失函数，具体用于计算预测误差，一旦整个卷积神经网络100的前向传播(如图2由110至140的传播为前向传播)完成，反向传播(如图2由140至110的传播为反向传播)就会开始更新前面提到的各层的权重值以及偏差，以减少卷积神经网络100的损失及卷积神经网络100通过输出层输出的结果和理想结果之间的误差。

[0159] 需要说明的是，如图2所示的卷积神经网络100仅作为一种卷积神经网络的示例，在具体的应用中，卷积神经网络还可以以其他网络模型的形式存在，例如，如图3所示的多个卷积层/池化层并行，将分别提取的特征均输入给全神经网络层130进行处理。

[0160] 下面介绍本申请实施例提供的一种芯片硬件结构。

[0161] 图4为本申请实施例提供的一种芯片硬件结构，该芯片包括神经网络处理器40。该芯片可以被设置在如图1所示的执行设备110中，用以完成计算模块111的计算工作。该芯片也可以被设置在如图1所示的训练设备120中，用以完成训练设备120的训练工作并输出目标模型/规则101。如图2所示的卷积神经网络中各层的算法均可在如图4所示的芯片中得以实现。

[0162] 神经网络处理器40可以是神经网络处理器(neural-network processing unit，NPU)，张量处理器(tensor processing unit，TPU)，或者图形处理器(graphics processing unit，GPU)等一切适合用于大规模异或运算处理的处理器。以NPU为例：神经网络处理器NPU40作为协处理器挂载到主中央处理器(central processing unit，CPU)(host CPU)上，由主CPU分配任务。NPU的核心部分为运算电路403，控制器404控制运算电路403提取存储器(权重存储器或输入存储器)中的数据并进行运算。

[0163] 在一些实现中，运算电路403内部包括多个处理单元(process engine，PE)。在一些实现中，运算电路403是二维脉动阵列。运算电路403还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中，运算电路403是通用的矩阵处理器。

[0164] 举例来说，假设有输入矩阵A，权重矩阵B，输出矩阵C。运算电路从权重存储器402中取矩阵B相应的数据，并缓存在运算电路中每一个PE上。运算电路从输入存储器401中取矩阵A数据与矩阵B进行矩阵运算，得到的矩阵的部分结果或最终结果，保存在累加器408中。

[0165] 向量计算单元407可以对运算电路的输出做进一步处理，如向量乘，向量加，指数运算，对数运算，大小比较等等。例如，向量计算单元407可以用于神经网络中非卷积/非FC层的网络计算，如池化(pooling)，批归一化(batch normalization)，局部响应归一化(local response normalization)等。

[0166] 在一些实现种，向量计算单元能407将经处理的输出的向量存储到统一缓存器406。例如，向量计算单元407可以将非线性函数应用到运算电路403的输出，例如累加值的向量，用以生成激活值。在一些实现中，向量计算单元407生成归一化的值、合并值，或二者均有。在一些实现中，处理过的输出的向量能够用作到运算电路403的激活输入，例如用于

在神经网络中的后续层中的使用。

[0167] 统一存储器406用于存放输入数据以及输出数据。

[0168] 权重数据直接通过存储单元访问控制器405(direct memory access controller,DMAC)将外部存储器中的输入数据搬运到输入存储器401和/或统一存储器406、将外部存储器中的权重数据存入权重存储器402,以及将统一存储器506中的数据存入外部存储器。

[0169] 总线接口单元(bus interface unit,BIU)410,用于通过总线实现主CPU、DMAC和取指存储器409之间进行交互。

[0170] 与控制器404连接的取指存储器(instruction fetch buffer)409,用于存储控制器404使用的指令。

[0171] 控制器404,用于调用指存储器409中缓存的指令,实现控制该运算加速器的工作过程。

[0172] 一般地,统一存储器406,输入存储器401,权重存储器402以及取指存储器409均为片上(On-Chip)存储器,外部存储器为该NPU外部的存储器,该外部存储器可以为双倍数据率同步动态随机存储器(double data rate synchronous dynamic random access memory,简称DDR SDRAM)、高带宽存储器(high bandwidth memory,HBM)或其他可读可写的存储器。

[0173] 其中,图2或图3所示的卷积神经网络中各层的运算可以由运算电路403或向量计算单元407执行。

[0174] 首先,先对本申请实施例提供的语音处理方法所适用的应用场景进行描述。该语音处理方法可以应用于需要修改语音内容的场景,例如:用户录制短视频、老师在录制授课语音等场景。该语音处理方法可以适用于例如手机、计算机、可发声的拆戴式终端上的智能语音助手、智能音响等具有语音编辑功能的应用程序、软件或语音处理设备上。

[0175] 其中,语音处理设备是一种用于服务用户的终端设备,或者云端设备。终端设备可以包括头戴显示设备(head mount display,HMD)、该头戴显示设备可以是虚拟现实(virtual reality,VR)盒子与终端的组合,VR一体机,个人计算机(personal computer,PC),增强现实(augmented reality,AR)设备,混合现实(mixed reality,MR)设备等,该终端设备还可以包括蜂窝电话(cellular phone)、智能电话(smart phone)、个人数字助理(personal digital assistant,PDA)、平板型电脑、膝上型电脑(laptop computer)、个人电脑(personal computer,PC)、车载终端等,具体此处不做限定。

[0176] 下面结合附图对本申请实施例的神经网络、预测网络的训练方法、语音处理方法进行详细的介绍。

[0177] 本申请实施例中的神经网络与预测网络可以是单独的两个网络,也可以是一个多任务的神经网络,其中一个任务是输出时长,另外一个任务是输出语音特征。

[0178] 其次,结合图5对本申请实施例的神经网络的训练方法进行详细介绍。图5所示的训练方法可以由神经网络的训练装置来执行,该神经网络的训练装置可以是云服务设备,也可以是终端设备,例如,电脑、服务器等运算能力足以用来执行神经网络的训练方法的装置,也可以是由云服务设备和终端设备构成的系统。示例性地,训练方法可以由图1中的训练设备120、图4中的神经网络处理器40执行。

[0179] 可选地，训练方法可以由CPU处理，也可以由CPU和GPU共同处理，也可以不用GPU，而使用其他适合用于神经网络计算的处理器，本申请不做限制。

[0180] 图5所示的训练方法包括步骤501与步骤502。下面对步骤501与步骤502进行详细说明。

[0181] 首先，先对预测网络的训练过程进行简单描述。本申请实施例中的预测网络可以是transformer网络、RNN、CNN等，具体此处不做限定。预测网络在训练阶段，输入是训练文本的向量，输出是训练文本中各个音素的时长、音高特征或者语音特征。再不断缩小预测网络输出的训练文本中各个音素的时长、音高特征或者语音特征与训练文本对应训练语音的实际时长、实际音高特征或者实际语音特征之间的差异，进而得到训练好的预测网络。

[0182] 步骤501，获取训练数据。

[0183] 本申请实施例中的训练数据包括训练语音，或者包括训练语音以及与训练语音对应的训练文本。如果训练数据不包括训练文本，则可以通过识别训练语音的方式获取训练文本。

[0184] 可选地，若发音者(或者用户)的数量为多个，为了后续预测的语音特征正确，训练数据中的训练语音特征还可以包括用户标识，或者包括训练语音的声纹特征，或者包括用于标识训练语音的声纹特征的向量。

[0185] 可选地，训练数据还可以包括训练语音中各个音素的起止时长信息。

[0186] 本申请实施例中获取训练数据可以是通过直接录制发声对象发声的方式获取，也可以是通过用户输入音频信息、视频信息的方式获取，还可以是通过接收采集设备发送的方式获取，在实际应用中，还有其他方式获取训练数据，对于训练数据的获取方式具体此处不做限定。

[0187] 步骤502，以训练数据作为神经网络的输入，以损失函数的值小于阈值为目标对神经网络进行训练，得到训练好的神经网络。

[0188] 可选地，训练数据可以进行一些预处理，例如上述所描述的如果训练数据包括训练语音，可以识别训练语音的方式获取训练文本，并将训练文本用音素表示输入神经网络。

[0189] 在训练过程中，可以将整个训练文本当做目标编辑文本，并作为输入，以减小损失函数的值为目标对神经网络进行训练，也就是不断缩小神经网络输出的语音特征与训练语音对应的实际语音特征之间的差异。该训练过程可以理解为预测任务。损失函数可以理解为预测任务对应的损失函数。

[0190] 本申请实施例中的神经网络具体可以是注意力机制模型，例如：transformer、tacotron2等。其中，注意力机制模型包括编码器-解码器，编码器或解码器的结构可以是循环神经网络、长短期记忆网络(long short-term memory，LSTM)等。

[0191] 本申请实施例中的神经网络包括编码器(encoder)与解码器(decoder)，编码器与解码器的结构类型可以是RNN、LSTM等，具体此处不做限定。编码器的作用是将训练文本编码为文本向量(以音素为单位的向量表示，每个输入对应一个向量)，解码器的作用是根据文本向量得到文本对应的语音特征。解码器在训练过程中，每步的计算以上一步所对应的真实语音特征作为条件进行计算。

[0192] 进一步的，为了保证前后语音的连贯，可以使用预测网络对文本向量对应的语音时长进行修正。即可以理解为根据训练语音中各个音素的时长对文本向量进行上采样(也

可以理解为是对向量的帧数进行扩展)，以得到对应帧数的向量。解码器的作用是根据上述对应帧数的向量得到文本对应的语音特征。

[0193] 可选地，上述的解码器可以是单向解码器，也可以是双向解码器(即两个方向并行)，具体此处不做限定。其中，两个方向是指训练文本的方向，也可以理解为是训练文本对应的向量的方向，还可以理解为是训练文本的正序或者反序，一个方向是训练文本的一侧指向训练文本的另一侧，另一个方向是训练文本的另一侧指向训练文本的一侧。

[0194] 示例性的，若训练文本为：“中午吃饭了没”，则第一方向或正序可以是从“中”到“没”的方向，第二方向或反序可以是从“没”到“中”的方向。

[0195] 若解码器是双向解码器，则两个方向(或者正反序)的解码器并行训练，且在训练过程中各自独立计算，不存在结果依赖。当然，如果预测网络与神经网络为一个多任务的网络，预测网络可以称为预测模块，则解码器可以根据训练文本对应的真实时长信息修正神经网络输出的语音特征。

[0196] 示例性的，以歌声编辑为例，进行模型训练时的输入可以为原始歌声音频、对应歌词文本(以音素为单位表示)根据原始歌声音频得到各音素在原始音频中的时长信息、唱歌人声纹特征，帧级的Pitch信息等，例如可以通过预训练的其他模型或工具(如歌声歌词对齐工具，Singer声纹提取工具，以及Pitch提取算法等)得到。输出可以为训练好的声学模型，训练目标为最小化预测出的歌声特征同歌声语音特征间的误差。

[0197] 在训练样本的数据准备上，可以基于歌声合成训练数据集，分别模拟“插入，删除和替换”操作场景构建对应的训练数据样本。

[0198] 训练过程：

[0199] Stage1：先使用ground-truth歌词和音频以及Pitch和duration数据，训练一个歌声合成模型，从而得到训练好的文本编码模块和音频特征解码模块；

[0200] Stage2：固定文本编码模块和音频特征解码模块，使用模拟编辑操作训练数据集训练时长规整模块和Pitch预测模块；

[0201] Stage3：端到端训练，使用所有的训练数据finetune整个模型。

[0202] 本申请实施例中的神经网络的架构可以参阅图6。其中，神经网络包括编码器与解码器。可选地，神经网络还可以包括预测模块与上采样模块。预测模块具体用于实现上述预测网络的功能，上采样模块具体用于实现上述根据训练语音中各个音素的时长对文本向量进行上采样的过程，具体此处不再赘述。

[0203] 需要说明的是，训练过程也可以不采用前述训练方法而采用其他训练方法，此处不做限定。

[0204] 下面结合附图对本申请实施例的语音处理方法进行详细的介绍。

[0205] 首先，本申请实施例提供的语音处理方法可以应用于替换场景、插入场景或删除场景。上述场景可以理解为是对原始文本对应的原始语音进行替换、插入、删除等得到目标语音，实现目标语音与原始语音的听感类似和/或提升目标语音的流畅度。其中，原始语音可以认为是包括待修改的语音，目标语音为用户想修正原始语音后得到的语音。

[0206] 为了方便理解，下面对上述场景的几种举例进行描述：

[0207] 一、对于替换场景。

[0208] 原始文本为“今天深圳天气很好”，目标文本为“今天广州天气很好”。其中，重叠文

本为“今天天气很好”。原始文本中的非重叠文本为“深圳”，目标文本中的非重叠文本为“广州”。目标文本包括第一文本与第二文本，第一文本为重叠文本或重叠文本中的部分文本。第二文本为目标文本中除了第一文本以外的文本。例如：若第一文本为“今天天气很好”，则第二文本为“广州”。若第一文本为“今气很好”，则第二文本为“天广州天”。

[0209] 二、对于插入场景。

[0210] 原始文本为“今天深圳天气很好”，目标文本为“今天上午深圳天气很好”。其中，重叠文本为“今天深圳天气很好”。目标文本中的非重叠文本为“上午”。为了实现目标语音前后的连贯，可以将该插入场景看作为将原始语音中的“天深”替换为“天上午深”的替换场景。即第一文本为“今圳天气很好”，第二文本为“天上午深”。

[0211] 三、对于删除场景。

[0212] 原始文本为“今天深圳天气很好”，目标文本为“今天天气很好”。其中，重叠文本为“今天天气很好”。原始文本中的非重叠文本为“深圳”。为了实现目标语音前后的连贯，可以将该删除场景看作为将原始语音中的“天深圳天”替换为“天天”的替换场景。即第一文本为“今气很好”，第二文本为“天天”。

[0213] 可选地，上述几种场景只是举例，在实际应用中，还有其他场景，具体此处不做限定。

[0214] 由于上述的删除场景与插入场景都可以用替换场景进行代替，下面仅以替换场景为例对本申请实施例提供的语音处理方法进行描述。本申请实施例提供的语音处理方法可以由终端设备或云端设备单独执行，也可以由终端设备与云端设备共同完成，下面分别描述：

[0215] 实施例一：终端设备或者云端设备单独执行该语音处理方法。

[0216] 请参阅图7a，本申请实施例提供的语音处理方法一个实施例，该方法可以由语音处理设备执行，也可以由语音处理设备的部件(例如处理器、芯片、或芯片系统等)执行，该语音处理设备可以是终端设备或云端设备，该实施例包括步骤701至步骤704。

[0217] 步骤701，获取原始语音与第二文本。

[0218] 本申请实施例中，语音处理设备可以直接获取原始语音、原始文本与第二文本。也可以先获取原始语音与第二文本，在识别原始语音得到与原始语音对应的原始文本。其中，第二文本为目标文本中除了第一文本以外的文本，且原始文本与目标文本含有第一文本。第一文本可以理解为是原始文本与目标文本的重叠文本中的部分或全部文本。

[0219] 在一种可能的实现中，所述原始语音的内容为用户的歌声，例如可以为用户清唱时录制的语音。

[0220] 本申请实施例中，语音处理设备获取第二文本的方式有多种，下面分别描述：

[0221] 第一种，语音处理设备可以通过其他设备或用户的输入直接获取第二文本。

[0222] 第二种，语音处理设备获取目标文本，并根据目标文本与原始语音对应的原始文本得到重叠文本，再根据重叠文本确定第二文本。具体可以是将原始文本与目标文本中的字符一一对比或者输入对比模型，确定原始文本与目标文本的重叠文本和/或非重叠文本。再根据重叠文本确定第一文本。其中，第一文本可以是重叠文本，也可以是重叠文本中的部分文本。

[0223] 本申请实施例中根据重叠文本确定第一文本的方式有多种，语音处理设备可以直

接确定重叠文本为第一文本，还可以根据预设规则确定重叠文本中的第一文本，也可以根据用户的操作确定重叠文本中的第一文本。其中，预设规则可以是去掉重叠内容中的N个字符后得到第一文本，N为正整数。

[0224] 可以理解的是，上述两种方式只是举例，在实际应用中，还有其他方式获取第二文本的方式，具体此处不做限定。

[0225] 另外，语音处理设备可以将原始文本与原始语音对齐，确定原始文本中各个音素在原始语音中的起止位置，可以获知原始文本中各个音素的时长。进而获取第一文本对应的音素，也即是获取第一文本在原始语音中对应的语音(即非编辑语音)。

[0226] 可选地，语音处理设备可以将原始文本与原始语音对齐采用的方式可以是采用强制对齐法，例如：蒙特利尔强制校准器(montreal forced aligner，MFA)、具有对齐功能的神经网络等对齐工具，具体此处不做限定。

[0227] 可选地，语音处理设备获取原始语音与原始文本之后，可以向用户展示用户界面，该用户界面包括原始语音以及原始文本。进一步的，用户通过用户界面对原始文本执行第一操作，语音处理设备响应用户的第一操作确定目标文本。其中，第一操作可以理解为是用户对原始文本的编辑，编辑具体可以是前述的替换、插入或删除等。

[0228] 示例性的，延续上述替换场景中的举例。原始文本为“今天深圳天气很好”，目标文本为“今天广州天气很好”。示例性的，以语音处理设备是手机为例进行描述。语音处理设备获取原始文本与原始语音之后，向用户展示如图8所示的界面，该界面包括原始文本与原始语音。如图9所示，用户可以对原始文本执行第一操作901，例如将“深圳”修改为“广州”等前述的插入、删除、替换操作，这里仅以替换为例进行描述。

[0229] 可选地，语音处理设备确定原始文本与目标文本的重叠文本后，向用户展示重叠文本，再根据用户的第二操作，从重叠文本中确定第一文本，进而确定第二文本。其中，第二操作可以是点击、拖拽、滑动等操作，具体此处不做限定。

[0230] 示例性的，延续上述举例，第二文本为“广州”，第一文本为“今天天气很好”，非编辑语音为第一文本在原始语音中的语音。假设一个文字对应2帧，原始文本对应的原始语音包括16帧，则非编辑语音相当于原始语音中的第1帧至第4帧以及第9帧至第16帧。可以理解的是，在实际应用中，文字与语音帧的对应关系不一定是上述举例的1比2，上述举例只是为了方便理解非编辑区域，原始文本对应的帧数具体此处不做限定。确定目标文本之后，语音处理设备可以显示如图10所示界面，该界面可以包括第二文本、目标文本、原始语音中的非编辑语音与编辑语音，其中，第二文本为“广州”，目标文本为“今天广州天气很好”，非编辑语音为“今天天气很好”对应的语音，编辑语音为“深圳”对应的语音。也可以理解为是，随着用户编辑的目标文本，进而语音处理设备基于目标文本、原始文本以及原始语音确定原始语音中的非编辑语音。

[0231] 可选地，语音处理设备接收用户发送的编辑请求，该编辑请求中包括原始语音与第二文本。可选地，编辑请求还包括原始文本和/或发音者标识。当然，该编辑请求也可以包括原始语音与目标文本。

[0232] 步骤702，根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息，预测所述第二文本的第二音高特征。

[0233] 在一种可能的实现中，所述目标文本的信息，包括：所述目标文本中各个音素的文

本嵌入(text embedding)。

[0234] 在一种可能的实现中,可以根据目标文本,通过文本编码模块(Text Encoder)得到目标文本中各个音素的文本嵌入。例如,目标文本可以被转成对应的音素序列(如“爱怎么可以不问对错”所对应的音素即为其拼音的声母和韵母序列),继而输入到Text Encoder转换成对应的以音素为单位的文本嵌入。Text Encoder的网络结构可示例性的为Tacotron 2模型。

[0235] 在一种可能的实现中,可以获取到非编辑语音中各个音素的帧数(也可以称之为时长),并根据所述非编辑语音中各个音素的帧数以及所述目标文本的信息,预测所述第二文本中各个音素的帧数。

[0236] 在一种可能的实现中,预测所述第二文本中各个音素的帧数所使用的神经网络可以如图7b所示(例如可以为基于掩码机制的融合原始真实时长的时长预测模型),其以Text Encoder的输出和原始真实时长(Reference Duration,也就是第一文本中各个音素的时长)以及对应掩码作为输入而预测每个待编辑音素(也就是第二文本中各个音素)的时长(即对应音频中的帧数)。

[0237] 在一种可能的实现中,在得到目标文本(包括第一文本和第二文本)中各个音素的帧数后,可以根据每个音素所预测的时长,将每个文本嵌入(text embedding)进行上采样,得到对应帧数的embedding结果(示例性的,若音素ai预测的时长为10帧,则可以将ai所对应的text embedding复制N份,N为大于1的正数,例如N为10)。

[0238] 应理解,可选的,在歌声编辑的场景中,歌声本身会遵循一定的曲谱,而曲谱也就规定了每个字的发音时长和Pitch等。因此,在歌声编辑时,对于无需编辑区域(非编辑语音),其对应的时长和音高Pitch的信息是无需预测的,直接得到准确的真实数值并使用即可。

[0239] 接下来给出一个对第二文本进行时长预测的示意:

[0240] 参照图7b,Reference Durations为原始歌声音频中各音素的真实时长,其中虚线框则为第二文本中各个音素的待预测时长(由于此时未知,则可以0代替);而Edit Mask则用已标记待预测的音素(其中Mask=0表示需预测);Embedding Layer将Reference durations同edit Mask进行融合(例如可以通过执行内积运算进行融合),其结果继而同Text Embedding以及Singer Embedding(提取的声纹特征)累加。其中,1个FFT Block可以为一个Transformer block,示例性的,可以使用4个(即N=4)个FFT block;最终,模型预测出Mask=0所对应音素的时长,并同其他未编辑音素时长一同作为输出。

[0241] 在一种可能的实现中,预测出的第二文本中各个音素的时长可以被用于进行音高特征预测中各个输入的上采样,例如,进行音高特征预测的输入可以包括文本嵌入,上采样前的每个文本嵌入对应于一个音素,上采样后的文本嵌入包括对应的音素的帧数的数量的文本嵌入。

[0242] 在一种可能的实现中,还可以会根据非编辑语音,得到非编辑语音的第二语音特征。所述第二语音特征可以携带有如下信息的至少一种:所述非编辑语音的部分语音帧或全部语音帧;所述非编辑语音的声纹特征;所述非编辑语音的音色特征;所述非编辑语音的韵律特征;以及,所述非编辑语音的节奏特征。

[0243] 本申请实施例中的语音特征可以用于表示语音的特征(例如:音色、韵律、情感或

节奏等),语音特征的表现形式有多种,可以是语音帧、序列、向量等,具体此处不做限定。另外,本申请实施例中的语音特征具体可以是通过前述的PLP、LPC、MFCC等方法从上述表现形式中提取的参数。

[0244] 可选地,从非编辑语音中选取至少一个语音帧作为第二语音特征。进一步的,为了第一语音特征更加结合了上下文的第二语音特征。至少一个语音帧对应的文本可以为第一文本中与第二文本相邻的文本。

[0245] 可选地,将非编辑语音通过编码模型编码得到目标序列,将该目标序列作为第二语音特征。其中,编码模型可以是CNN、RNN等,具体此处不做限定。

[0246] 另外,第二语音特征还可以携带有原始语音的声纹特征。其中,获取声纹特征的方式可以是直接获取,也可以是通过识别原始语音得到该声纹特征等。一方面,通过引入原始语音的声纹特征,使得后续生成的第一语音特征也携带有该原始语音的声纹特征,进而提升目标编辑语音与原始语音的相近程度。另一方面,在发音者(或者用户)的数量为多个的情况下,引入声纹特征可以提升后续预测的语音特征更加与原始语音的发音者的声纹相似。

[0247] 可选地,语音处理设备还可以获取原始语音的发音者标识,以便于在发音者为多个时,可以匹配相应发音者对应的语音,提升后续目标编辑语音与原始语音的相似度。

[0248] 下面仅以将语音帧作为语音特征(或者理解为是根据语音帧获取语音特征)为例进行描述。示例性的,延续上述举例,选择原始语音中的第1帧至第4帧以及第9帧至第16帧中的至少一帧作为第二语音特征。

[0249] 示例性的,第二语音特征为梅尔频谱特征。

[0250] 在一种可能的实现中,第二语音特征可以为向量的形式表达,在一种可能的实现中,预测出的第二文本中各个音素的时长可以被用于进行音高特征预测中各个输入的上采样,例如,进行音高特征预测的输入可以包括第二语音特征,上采样前的每个向量对应于一个音素,上采样后的文本嵌入包括对应的音素的帧数的数量的向量。

[0251] 在一种可能的实现中,可以根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息,预测所述第二文本的第二音高特征。

[0252] 在一种可能的实现中,非编辑语音的第一音高(pitch)特征可以通过现有的Pitch提取算法得到,本申请并不限定。

[0253] 在一种可能的实现中,可以根据所述非编辑语音的第一音高(pitch)特征、所述目标文本的信息以及所述非编辑语音的第二语音特征,通过神经网格来预测所述第二文本的第二音高特征。

[0254] 接下来介绍如何根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息,预测所述第二文本的第二音高特征:

[0255] 在一种可能的实现中,所述目标文本为将所述第二文本插入到所述第一文本得到的文本;或者,所述目标文本为将所述第一文本的第一部分文本删除得到的文本,所述第二文本为与所述第一部分文本相邻的文本;可以将所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息进行融合,以得到第一融合结果;将所述第一融合结果输入到第二神经网络,得到所述第二文本的第二音高特征。

[0256] 针对插入和删除操作:使用图7c所示的模型预测目标编辑音素的帧级别的pitch

特征。针对插入和删除操作的Pitch预测模型，其模型结构可以同图7b设置一致或者相似，区别仅在于此时的输入是帧级别(图7b中输入的是音素级别的)的从真实歌声中所提取的pitch值(图7b中输入的是时长信息)，其中待编辑区域的pitch如虚框标记，且其对应Edit Mask标记设置为0。

[0257] 在一种可能的实现中，所述目标文本为将所述第一文本中的第二部分文本替换为所述第二文本得到的；可以将所述非编辑语音的第一音高(pitch)特征输入到第三神经网络，得到初始音高特征，所述第一初始音高特征包括多个帧中每个帧的音高；将所述目标文本的信息输入到第四神经网络，得到所述第二文本的发音特征，所述发音特征用于指示所述初始音高特征包括的多个帧中各个帧是否发音；将所述初始音高特征和所述发音特征进行融合，以得到所述第二文本的第二音高特征。

[0258] 针对于替换操作(此处替换操作仅表示新编辑文本字数同被替换文本字数一致的情况，若不一致，则将替换操作分解成先删除后插入两个编辑操作)。由于替换的文本可能在发音上存在很大区别，所以为保障替换后前后歌声的连贯性，使用图7d所示的模型来预测新pitch：

[0259] 针对替换操作的Pitch预测模型。可以引入帧级别的发音/未发音(Voiced/Unvoicde，U/UV)预测来帮助Pitch的预测。示例性的，V/UV Predictor和F0 Predictor模块的设计可参照Fastspeech2中F0 predictor。

[0260] 在一种可能的实现中，输入的所述第一音高(pitch)特征可以包括所述非编辑语音的多帧中的每一帧的音高特征；相应的，输出的所述第二音高特征可以包括所述目标编辑语音的多帧中的每一帧的音高特征。

[0261] 步骤703，根据所述第二音高特征以及所述第二文本，通过神经网络得到所述第二文本对应的第一语音特征。

[0262] 在一种可能的实现中，可以将第二音高特征以及所述第二文本(例如第二文本的文本嵌入)进行融合(例如相加)，并将融合结果输入到神经网络中，以得到第二文本对应的第一语音特征。其中，第二文本对应的第一语音特征可以为梅尔频谱特征。

[0263] 在一种可能的实现中，可以根据所述非编辑语音的第一音高(pitch)特征、所述目标文本的信息以及所述非编辑语音的第二语音特征，关于第二语音特征的描述可以参照上述实施例中关于第二语音特征的描述，这里不再赘述。

[0264] 在一种可能的实现中，在获取第二语音特征之后，可以基于第二语音特征、第二文本通过神经网络得到第二文本对应的第一语音特征。该神经网络可以包括编码器与解码器。将第二文本输入编码器得到第二文本对应的第一向量，再基于第二语音特征通过解码器对第一向量进行解码得到第一语音特征。其中，第二语音特征可以与第一语音特征的韵律、音色和/或信噪比等相同或相近，韵律可以反映出发音者的情感状态或讲话形式等，韵律泛指语调、音调、重音强调、停顿或节奏等特征。

[0265] 可选地，编码器与解码器之间可以引入注意力机制，用于调整输入与输出之间数量的对应关系。

[0266] 可选地，在编码器编码过程中可以引入第二文本所在的目标文本，使得生成的第二文本的第一向量参考了目标文本，使得该第一向量描述的第二文本更加准确。即可以基于第二语音特征、目标文本、标记信息通过神经网络得到第二文本对应的第一语音特征。具

体可以是将目标文本与标记信息输入编码器得到第二文本对应的第一向量，再基于第二语音特征通过解码器对第一向量进行解码得到第一语音特征。该标记信息用于标记目标文本中的第二文本。

[0267] 本申请实施例中的解码器可以是单向解码器，也可以是双向解码器，下面分别描述。

[0268] 第一种，解码器是单向解码器。

[0269] 解码器基于第二语音特征从目标文本的第一方向计算第一向量或第二向量得到的语音帧作为第一语音特征。其中，第一方向为从目标文本的一侧指向目标文本的另一侧的方向。另外，该第一方向可以理解为是目标文本的正序或反序(相关描述可以参考前述图5所示实施例中关于正序反序的描述)。

[0270] 可选地，将第二语音特征与第一向量输入解码器得到第一语音特征。或者将第二语音特征与第二向量输入解码器得到第一语音特征。

[0271] 第二种，若第二文本在目标文本的中间区域，解码器可以是双向解码器(也可以理解为编码器包括第一编码器与第二编码器)。

[0272] 上述的第二文本在目标文本的中间区域，可以理解为第二文本并不在目标文本的两端。

[0273] 本申请实施例中的双向解码器有多种情况，下面分别描述：

[0274] 1、双向解码器从第一方向输出的第一语音特征为第二文本对应的语音特征，双向解码器从第二方向输出的第四语音特征为第二文本对应的语音特征。

[0275] 该种情况，可以理解为可以分别通过左右两侧(即正序反序)得到两种第二文本对应的完整语音特征，并根据两种语音特征得到第一语音特征。

[0276] 第一解码器基于第二语音特征从目标文本的第一方向计算第一向量或第二向量得到第二文本的第一语音特征(以下称为LR)。第二解码器基于第二语音特征从目标文本的第二方向计算第一向量或第二向量得到第二文本的第四语音特征(以下称为RL)。并根据第一语音特征与第四语音特征生成第一语音特征。其中，第一方向为从目标文本的一侧指向目标文本的另一侧的方向，第二方向与第一方向相反(或者理解为第二方向为从目标文本的另一侧指向目标文本的一侧方向)。第一方向可以是上述的正序，第二方向可以是上述的反序。

[0277] 对于双向解码器，第一编码器在第一方向解码第一向量或第二向量的第一帧时，可以将非编辑语音中与第二文本一侧(也可以称为左侧)相邻的语音帧作为条件进行解码得到N帧LR。第二编码器在第二方向解码第一向量或第二向量的第一帧时，可以将非编辑语音中与第二文本另一侧(也可以称为右侧)相邻的语音帧作为条件进行解码得到N帧RL。可选地，双向解码器的结构可以参考图11。获取N帧LR与N帧RL之后，可以将LR与RL中差值小于阈值的帧作为过渡帧(位置为m，$m<n$，)，或者将LR与RL中差值最小的帧作为过渡帧。则第一语音特征的N帧可以包括LR中的前m帧与RL中的后n-m帧，或者第一语音特征的N帧包括LR中的前n-m帧与RL中的后m帧。其中，LR与RL的差值可以理解为是向量与向量之间的距离。另外，若前述步骤701中获取了发音者标识，则本步骤中的第一向量或第二向量还可以包括用于标识发音者的第三向量。也可以理解为第三向量用于标识原始语音的声纹特征。

[0278] 示例性的，延续上述举例，假设第一编码器得到“广州”对应的LR帧包括$LR_1$、$LR_2$、

$LR_3$、$LR_4$。第二编码器得到“广州”对应的RL帧包括$RL_1$、$RL_2$、$RL_3$、$RL_4$。且$LR_2$与$RL_2$差值最小，则将$LR_1$、$LR_2$、$RL_3$、$RL_4$或者$LR_1$、$RL_2$、$RL_3$、$RL_4$作为第一语音特征。

[0279] 2、双向解码器从第一方向输出的第一语音特征为第二文本中第三文本对应的语音特征，双向解码器从第二方向输出的第四语音特征为第二文本中第四文本对应的语音特征。

[0280] 该种情况，可以理解为可以分别通过左右两侧(即正序反序)得到第二文本对应的部分语音特征，并根据两个部分语音特征得到完整的第一语音特征。即从正序的方向上取一部分语音特征，从反序的方向上取另一部分语音特征，并拼接一部分语音特征与另一部分语音特征得到整体的语音特征。

[0281] 示例性的，延续上述举例，假设第一编码器得到第三文本(“广”)对应的LR帧包括$LR_1$与$LR_2$。第二编码器得到第四文本(“州”)对应的RL帧包括$RL_3$与$RL_4$。则拼接$LR_1$、$LR_2$、$RL_3$、$RL_4$得到第一语音特征。

[0282] 可以理解的是，上述两种方式只是举例，在实际应用中，还有其他方式获取第一语音特征，具体此处不做限定。

[0283] 步骤704，根据所述第一语音特征，生成所述第二文本对应的目标编辑语音。

[0284] 在一种可能的实现中，在获取第一语音特征之后，可以根据声码器将第一语音特征转换为第二文本对应的目标编辑语音。其中，声码器可以是传统声码器(例如Griffin-lim算法)，也可以是神经网络声码器(如使用音频训练数据预训练好的Melgan，或Hifigan等)等，具体此处不做限定。

[0285] 示例性的，延续上述举例，“广州”对应的目标编辑语音如图12所示。

[0286] 步骤705，获取第二文本在目标文本中的位置。本步骤是可选地。

[0287] 可选地，如果步骤701中获取的是原始语音与第二文本，则获取第二文本在目标文本中的位置。

[0288] 可选地，如果步骤701中已获取目标文本，则可以通过前述步骤701中的对齐技术对齐原始语音与原始文本确定原始文本中各个音素在原始语音中的起止位置。并根据各音素的起止位置确定第二文本在目标文本中的位置。

[0289] 步骤706，基于位置拼接目标编辑语音与非编辑语音生成与目标文本对应的目标语音。本步骤是可选地。

[0290] 本申请实施例中的位置用于拼接非编辑语音与目标编辑语音，该位置可以是第二文本在目标文本中的位置，也可以是第一文本在目标文本中的位置，还可以是非编辑语音在原始语音中的位置，还可以是编辑语音在原始语音中的位置。

[0291] 可选地，获取第二文本在目标文本中的位置之后，可以通过前述步骤701中的对齐技术对齐原始语音与原始文本确定原始文本中各个音素在原始语音中的起止位置。并根据第一文本在原始文本中的位置，确定原始语音中的非编辑语音或编辑语音位置。进而语音处理设备基于位置拼接目标编辑语音与非编辑语音得到目标语音。即将第二文本对应的目标语音替换原始语音中的编辑区域得到目标语音。

[0292] 示例性的，延续上述举例，非编辑语音相当于原始语音中的第1帧至第4帧以及第9帧至第16帧。目标编辑语音为$LR_1$、$LR_2$、$RL_3$、$RL_4$或者$LR_1$、$RL_2$、$RL_3$、$RL_4$。拼接目标编辑语音与非编辑语音，可以理解为是将得到的四帧替换原始语音中的第5帧至第8帧，进而得到目标语

音。即将“广州”对应的语音替换原始语音中“深圳”对应的语音，进而得到目标文本：“今天广州天气很好”对应的目标语音。“今天广州天气很好”对应的目标语音如图12所示。

[0293] 可选地，语音处理设备在获取目标编辑语音或目标语音之后，对目标编辑语音或目标语音进行播放。

[0294] 一种可能实现的方式中，本申请实施例提供的语音处理方法包括步骤701至步骤704。另一种可能实现的方式中，本申请实施例提供的语音处理方法包括步骤701至步骤705。另一种可能实现的方式中，本申请实施例提供的语音处理方法包括步骤701至步骤706。另外，本申请实施例中图7a所示的各个步骤不限定时序关系。例如：上述方法中的步骤705也可以在步骤704之后，也可以在步骤701之前，还可以与步骤701共同执行。

[0295] 本申请实施例提供了一种语音处理方法，所述方法包括：获取原始语音以及第二文本，所述第二文本为目标文本中除了第一文本以外的文本，所述目标文本与所述原始语音对应的原始文本都包括所述第一文本，所述第一文本在所述原始语音中对应的语音为非编辑语音；根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息，预测所述第二文本的第二音高特征；根据所述第二音高特征以及所述第二文本，通过神经网络得到所述第二文本对应的第一语音特征；根据所述第一语音特征，生成所述第二文本对应的目标编辑语音。本申请通过预测第二文本(待编辑文本)的音高特征，根据音高特征生成第二文本的第一语音特征，并基于第一语音特征生成第二文本对应目标编辑语音，使得歌声编辑前后的语音的音高特征相似，进而实现目标编辑语音的听感与原始语音的听感目标编辑语音的听感与原始语音的听感类似。

[0296] 接下来结合一个示意介绍本申请实施例中的语音处理方法：

[0297] 以歌声编辑的场景为例，分别以原始待编辑歌声W(其中语音内容为S“爱可以不问对错”)，和以下三条不同目标语音为例：

[0298] 编辑请求Q1：其目标语音为W1(语音内容对应文本T1为“爱怎么可以不问对错”)，

[0299] 编辑请求Q2：其目标语音为W2(语音内容对应文本T2为“爱不问对错”)，

[0300] 编辑请求Q3：其目标语音为W3(语音内容对应文本T2为“爱怎么不问对错”)

[0301] 步骤S1：接收用户“语音编辑”请求；

[0302] 该请求至少包括原始待编辑语音W，原始歌词文本S，目标文本T(T1或T2或T3)等数据，预操作包括：对比原始文本S和目标文本，确定当前编辑请求的编辑类型：即对于Q1，Q2和Q3，可确定得到它们分别为插入，删除和替换操作；从W中提取每帧的音频特征，Pitch特征；W经声纹模型提取出Singer embedding；将S和目标文本T*转换成音素的表示形式，比如T2，其音素序列为[ai4 b u2 w en4 d ui4 c cuo4]；根据W和S，提取S中每个音素所对应的时长(即帧数)；根据操作类型确定Mask区域，对于Q1，其为插入操作(插入词“怎么”)则其目标Mask音素为“怎么”所对应的音素，即Q1最终的目标文本音素为[ai4z en3 m e5 k e2 y i3 b u2 w en4 d ui4 c cuo4]；(其中红色表示被Mask的音素)，对于Q2，其为删除操作(删除词“可以”)，则其目标音素为S中原本同“可以”相邻词的音素；即Q2最终的目标文本音素为[ai4 b u2 w en4 d ui4 c cuo4]；(其中红色表示被Mask的音素)；对于Q3，其为替换操作(将“可以”替换成“怎么”)，所以其目标文本音素为[ai4 z en3 m e5 b u2 w en4 d ui4 c cuo4]；(其中红色表示被Mask的音素)；

[0303] 步骤S2：S1中所得的目标文本音素经文本编码模块生成文本特征，即Phoneme-

level Text Embedding;

[0304] 步骤S3:经时长规整模块预测出目标文本中各音素的时长信息;该步骤可通过以下子步骤完成:

[0305] 根据音素的Mask标记生成Mask向量和参考时长向量:即对于非Mask音素,其参考时长即为S1步骤所提取的真实时长,否则则设为0;对于非Mask音素,Mask向量中对应位置设置为1,否则设置为0;

[0306] 将Text Embedding,Singer Embedding参考时长向量以及Mask向量作为输入,使用如Figure2-2所示的时长预测模块预测出Mask音素所对应的时长

[0307] 根据各音素所对应的时长,将各音素的Embedding向上采样(即若音素A的时长为10,则将A的Embedding复制10份),从而生成Frame-level Text Embedding;

[0308] 步骤S4:经Pitch预测模块预测出各帧的Pitch值,该步骤可通过以下子步骤完成:

[0309] 对于Q1和Q2,使用Figure2-3所示的模型预测出Mask音素所对应的帧的pitch:

[0310] 其中,对于非Mask音素,其参考pitch即为S1中提取的真实Pitch,且其在Mask向量对应位置上标记为1,;对于Mask音素,其对应帧上的pitch设置为0,Mask设置为0;预测出Mask音素所对应的Frame-level pitch。

[0311] 对于替换操作Q3,则使用Figure2-4所示的模型预测出Mask音素的Frame-level Pitch;

[0312] 步骤S5:将Frame-Level text Embedding和Pitch加到一起输入到音频特征解码模块,预测出新的Mask音素所对应的音频特征帧。

[0313] 应理解,若一个编辑请求中涉及多个编辑操作,则可以按照从左至右的处理顺序一一使用如上所述的流程进行编辑。另一方面,一个替换操作也可以通过“先删除后插入”两个操作来实现。

[0314] 上面对终端设备或云端设备单独实施的语音处理方法进行了描述,下面对终端设备与云端设备共同执行的语音处理方法进行描述。

[0315] 实施例二:终端设备与云端设备共同执行语音处理方法。

[0316] 请参阅图13,本申请实施例提供的语音处理方法一个实施例,该方法可以由终端设备与云端设备共同执行,也可以由终端设备的部件(例如处理器、芯片、或芯片系统等)与云端设备的部件(例如处理器、芯片、或芯片系统等)执行,该实施例包括步骤1301至步骤1306。

[0317] 步骤1301,终端设备获取原始语音与第二文本。

[0318] 本实施例中终端设备执行的步骤1301与前述图7a所示实施例中语音处理设备执行的步骤701类似,此处不再赘述。

[0319] 步骤1302,终端设备向云端设备发送原始语音与第二文本。

[0320] 终端设备获取原始语音与第二文本之后,可以向云端设备发送原始语音与第二文本。

[0321] 可选地,若步骤1301中,终端设备获取的是原始语音与目标文本,则终端设备向云端设备发送原始语音与目标文本。

[0322] 步骤1303,云端设备基于原始语音与第二文本获取非编辑语音。

[0323] 本实施例中云端设备执行的步骤1303与前述图7a所示实施例中语音处理设备执

行的步骤701中确定非编辑语音的描述类似，此处不再赘述。

[0324] 步骤1304，云端设备基于非编辑语音的第一音高特征和目标文本的信息，获取第二文本的第二音高特征。

[0325] 本实施例中云端设备执行的步骤1303与前述图7a所示实施例中语音处理设备执行的步骤702中确定非编辑语音的描述类似，此处不再赘述。

[0326] 步骤1305，云端设备基于第二音高特征、第二文本通过神经网络得到第二文本对应的第一语音特征。

[0327] 步骤1306，云端设备基于第一语音特征生成与第二文本对应的目标编辑语音。

[0328] 本实施例中云端设备执行的步骤1304至步骤1306与前述图7a所示实施例中语音处理设备执行的步骤702至步骤704类似，此处不再赘述。

[0329] 步骤1307，云端设备向终端设备发送目标编辑语音。本步骤是可选地。

[0330] 可选地，云端设备获取目标编辑语音之后，可以向终端设备发送目标编辑语音。

[0331] 步骤1308，终端设备或云端设备获取第二文本在目标文本中的位置。本步骤是可选地。

[0332] 步骤1309，终端设备或云端设备基于位置拼接目标编辑语音与非编辑语音生成与目标文本对应的目标语音。本步骤是可选地。本步骤是可选地。

[0333] 本实施例中的步骤1308、步骤1309与前述图7a所示实施例中语音处理设备执行的步骤705至步骤706类似，此处不再赘述。本实施例中的步骤1308、步骤1309可以由终端设备或云端设备执行。

[0334] 步骤1310，云端设备向终端设备发送目标语音。本步骤是可选地。

[0335] 可选地，若步骤1308与步骤1309由云端设备执行，则云端设备获取目标语音后，向终端设备发送目标语音。若步骤1308与步骤1309由终端设备执行，则可以不执行本步骤。

[0336] 可选地，终端设备在获取目标编辑语音或目标语音之后，对目标编辑语音或目标语音进行播放。

[0337] 一种可能实现的方式中，本申请实施例提供的语音处理方法可以包括：云端设备生成目标编辑语音，并向终端设备发送目标编辑语音，即该方法包括步骤1301至步骤1307。另一种可能实现的方式中，本申请实施例提供的语音处理方法可以包括：云端设备生成目标编辑语音，并根据目标编辑语音与非编辑语音生成目标语音，向终端设备发送目标语音。即该方法包括步骤1301至步骤1306、步骤1308至步骤1310。另一种可能实现的方式中，本申请实施例提供的语音处理方法可以包括：云端设备生成目标编辑语音，向终端设备发送目标编辑语音。终端设备在根据目标编辑语音与非编辑语音生成目标语音。即该方法包括步骤1301至步骤1309。

[0338] 本申请实施例中，一方面可以通过云端设备与终端设备的交互，由云端设备进行复杂的计算得到目标编辑语音或目标语音并返给终端设备，可以减少终端设备的算力与存储空间。另一方面，可以根据原始语音中非编辑区域的语音特征生成修改文本对应的目标编辑语音，进而与非编辑语音生成目标文本对应的目标语音。另一方面，用户可以通过对原始文本中的文本进行修改，得到修改文本(即第二文本)对应的目标编辑语音。提升用户基于文本进行语音编辑的编辑体验。另一方面，生成目标语音时，并未修改非编辑语音，且目标编辑语音的音高特征与非编辑语音的音高特征类似，使得用户在听原始语音与目标语音

时,很难听出原始语音与目标语音在语音特征上的差别。

[0339] 上面对本申请实施例中的语音处理方法进行了描述,下面对本申请实施例中的语音处理设备进行描述,请参阅图14,本申请实施例中语音处理设备的一个实施例包括:

[0340] 获取模块1401,用于获取原始语音以及第二文本,所述第二文本为目标文本中除了第一文本以外的文本,所述目标文本与所述原始语音对应的原始文本都包括所述第一文本,所述第一文本在所述原始语音中对应的语音为非编辑语音;

[0341] 其中,关于获取模块1401的具体描述可以参照上述实施例中步骤701的描述,这里不再赘述。

[0342] 音高预测模块1402,用于根据所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息,预测所述第二文本的第二音高特征;

[0343] 其中,关于音高预测模块1402的具体描述可以参照上述实施例中步骤702的描述,这里不再赘述。

[0344] 生成模块1403,用于根据所述第二音高特征以及所述第二文本,通过神经网络得到所述第二文本对应的第一语音特征;

[0345] 根据所述第一语音特征,生成所述第二文本对应的目标编辑语音。

[0346] 其中,关于生成模块1403的具体描述可以参照上述实施例中步骤703和704的描述,这里不再赘述。

[0347] 在一种可能的实现中,所述原始语音的内容为用户的歌声。

[0348] 在一种可能的实现中,所述根据所述非编辑语音的第一音高(pitch)特征以及所述第二文本包括:

[0349] 根据所述非编辑语音的第一音高(pitch)特征、所述目标文本的信息以及所述非编辑语音的第二语音特征;所述第二语音特征携带有如下信息的至少一种:

[0350] 所述非编辑语音的部分语音帧或全部语音帧;

[0351] 所述非编辑语音的声纹特征;

[0352] 所述非编辑语音的音色特征;

[0353] 所述非编辑语音的韵律特征;以及,

[0354] 所述非编辑语音的节奏特征。

[0355] 在一种可能的实现中,所述目标文本的信息,包括:所述目标文本中各个音素的文本嵌入(text embedding)。

[0356] 在一种可能的实现中,所述目标文本为将所述第二文本插入到所述第一文本得到的文本;或者,所述目标文本为将所述第一文本的第一部分文本删除得到的文本,所述第二文本为与所述第一部分文本相邻的文本;

[0357] 所述音高预测模块,具体用于:

[0358] 将所述非编辑语音的第一音高(pitch)特征以及所述目标文本的信息进行融合,以得到第一融合结果;

[0359] 将所述第一融合结果输入到第二神经网络,得到所述第二文本的第二音高特征。

[0360] 在一种可能的实现中,所述目标文本为将所述第一文本中的第二部分文本替换为所述第二文本得到的;

[0361] 所述音高预测模块,具体用于:

[0362] 将所述非编辑语音的第一音高(pitch)特征输入到第三神经网络,得到初始音高特征,所述第一初始音高特征包括多个帧中每个帧的音高;

[0363] 将所述目标文本的信息输入到第四神经网络,得到所述第二文本的发音特征,所述发音特征用于指示所述初始音高特征包括的多个帧中各个帧是否发音;

[0364] 将所述初始音高特征和所述发音特征进行融合,以得到所述第二文本的第二音高特征。

[0365] 在一种可能的实现中,所述装置还包括:

[0366] 时长预测模块,用于根据所述非编辑语音中各个音素的帧数以及所述目标文本的信息,预测所述第二文本中各个音素的帧数。

[0367] 在一种可能的实现中,所述第一音高(pitch)特征,包括:所述非编辑语音的多帧中的每一帧的音高特征;

[0368] 所述第二音高特征,包括:所述目标编辑语音的多帧中的每一帧的音高特征。

[0369] 在一种可能的实现中,所述时长预测模块,具体用于:

[0370] 根据所述非编辑语音中各个音素的帧数、所述目标文本的信息以及所述非编辑语音的第二语音特征。

[0371] 在一种可能的实现中,所述获取模块还用于:

[0372] 获取所述第二文本在所述目标文本中的位置;

[0373] 所述生成模块,还用于基于所述位置拼接所述目标编辑语音与所述非编辑语音得到所述目标文本对应的目标语音。

[0374] 请参阅图15,本申请实施例提供了另一种语音处理设备,为了便于说明,仅示出了与本申请实施例相关的部分,具体技术细节未揭示的,请参照本申请实施例方法部分。该语音处理设备可以为包括手机、平板电脑、个人数字助理(personal digital assistant,PDA)、销售终端设备(point of sales,POS)、车载电脑等任意终端设备,以语音处理设备为手机为例:

[0375] 图15示出的是与本申请实施例提供的语音处理设备相关的手机的部分结构的框图。参考图15,手机包括:射频(radio frequency,RF)电路1510、存储器1520、输入单元1530、显示单元1540、传感器1550、音频电路1560、无线保真(wireless fidelity,WiFi)模块1570、处理器1580、以及电源1590等部件。本领域技术人员可以理解,图15中示出的手机结构并不构成对手机的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件布置。

[0376] 下面结合图15对手机的各个构成部件进行具体的介绍:

[0377] RF电路1510可用于收发信息或通话过程中,信号的接收和发送,特别地,将基站的下行信息接收后,给处理器1580处理;另外,将设计上行的数据发送给基站。通常,RF电路1510包括但不限于天线、至少一个放大器、收发信机、耦合器、低噪声放大器(low noise amplifier,LNA)、双工器等。此外,RF电路1510还可以通过无线通信与网络和其他设备通信。上述无线通信可以使用任一通信标准或协议,包括但不限于全球移动通讯系统(global system of mobile communication,GSM)、通用分组无线服务(general packet radio service,GPRS)、码分多址(code division multiple access,CDMA)、宽带码分多址(wideband code division multiple access,WCDMA)、长期演进(long term evolution,

LTE)、电子邮件、短消息服务(short messaging service,SMS)等。

[0378] 存储器1520可用于存储软件程序以及模块,处理器1580通过运行存储在存储器1520的软件程序以及模块,从而执行手机的各种功能应用以及数据处理。存储器1520可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等;存储数据区可存储根据手机的使用所创建的数据(比如音频数据、电话本等)等。此外,存储器1520可以包括高速随机存取存储器,还可以包括非易失性存储器,例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

[0379] 输入单元1530可用于接收输入的数字或字符信息,以及产生与手机的用户设置以及功能控制有关的键信号输入。具体地,输入单元1530可包括触控面板1531以及其他输入设备1532。触控面板1531,也称为触摸屏,可收集用户在其上或附近的触摸操作(比如用户使用手指、触笔等任何适合的物体或附件在触控面板1531上或在触控面板1531附近的操作),并根据预先设定的程式驱动相应的连接装置。可选的,触控面板1531可包括触摸检测装置和触摸控制器两个部分。其中,触摸检测装置检测用户的触摸方位,并检测触摸操作带来的信号,将信号传送给触摸控制器;触摸控制器从触摸检测装置上接收触摸信息,并将它转换成触点坐标,再送给处理器1580,并能接收处理器1580发来的命令并加以执行。此外,可以采用电阻式、电容式、红外线以及表面声波等多种类型实现触控面板1531。除了触控面板1531,输入单元1530还可以包括其他输入设备1532。具体地,其他输入设备1532可以包括但不限于物理键盘、功能键(比如音量控制按键、开关按键等)、轨迹球、鼠标、操作杆等中的一种或多种。

[0380] 显示单元1540可用于显示由用户输入的信息或提供给用户的信息以及手机的各种菜单。显示单元1540可包括显示面板1541,可选的,可以采用液晶显示器(liquid crystal display,LCD)、有机发光二极管(organic light-emitting diode,OLED)等形式来配置显示面板1541。进一步的,触控面板1531可覆盖显示面板1541,当触控面板1531检测到在其上或附近的触摸操作后,传送给处理器1580以确定触摸事件的类型,随后处理器1580根据触摸事件的类型在显示面板1541上提供相应的视觉输出。虽然在图15中,触控面板1531与显示面板1541是作为两个独立的部件来实现手机的输入和输入功能,但是在某些实施例中,可以将触控面板1531与显示面板1541集成而实现手机的输入和输出功能。

[0381] 手机还可包括至少一种传感器1550,比如光传感器、运动传感器以及其他传感器。具体地,光传感器可包括环境光传感器及接近传感器,其中,环境光传感器可根据环境光线的明暗来调节显示面板1541的亮度,接近传感器可在手机移动到耳边时,关闭显示面板1541和/或背光。作为运动传感器的一种,加速计传感器可检测各个方向上(一般为三轴)加速度的大小,静止时可检测出重力的大小及方向,可用于识别手机姿态的应用(比如横竖屏切换、相关游戏、磁力计姿态校准)、振动识别相关功能(比如计步器、敲击)等;至于手机还可配置的陀螺仪、气压计、湿度计、温度计、红外线传感器等其他传感器,在此不再赘述。

[0382] 音频电路1560、扬声器1561,传声器1562可提供用户与手机之间的音频接口。音频电路1560可将接收到的音频数据转换后的电信号,传输到扬声器1561,由扬声器1561转换为声音信号输出;另一方面,传声器1562将收集的声音信号转换为电信号,由音频电路1560接收后转换为音频数据,再将音频数据输出处理器1580处理后,经RF电路1510以发送给比

如另一手机，或者将音频数据输出至存储器1520以便进一步处理。

[0383] WiFi属于短距离无线传输技术，手机通过WiFi模块1570可以帮助用户收发电子邮件、浏览网页和访问流式媒体等，它为用户提供了无线的宽带互联网访问。虽然图15示出了WiFi模块1570，但是可以理解的是，其并不属于手机的必须构成。

[0384] 处理器1580是手机的控制中心，利用各种接口和线路连接整个手机的各个部分，通过运行或执行存储在存储器1520内的软件程序和/或模块，以及调用存储在存储器1520内的数据，执行手机的各种功能和处理数据，从而对手机进行整体监控。可选的，处理器1580可包括一个或多个处理单元；优选的，处理器1580可集成应用处理器和调制解调处理器，其中，应用处理器主要处理操作系统、用户界面和应用程序等，调制解调处理器主要处理无线通信。可以理解的是，上述调制解调处理器也可以不集成到处理器1580中。

[0385] 手机还包括给各个部件供电的电源1590(比如电池)，优选的，电源可以通过电源管理系统与处理器1580逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。

[0386] 尽管未示出，手机还可以包括摄像头、蓝牙模块等，在此不再赘述。

[0387] 在本申请实施例中，该终端设备所包括的处理器1580可以执行前述图7a实施例中语音处理设备的功能，或者执行前述图13所示实施例中终端设备的功能，此处不再赘述。

[0388] 参阅图16，本申请提供的另一种语音处理设备的结构示意图。该语音处理设备可以是云端设备。该云端设备可以包括处理器1601、存储器1602和通信接口1603。该处理器1601、存储器1602和通信接口1603通过线路互联。其中，存储器1602中存储有程序指令和数据。

[0389] 存储器1602中存储了前述图7a对应的实施方式中，由语音处理设备执行的步骤对应的程序指令以及数据。或者存储了前述图13对应的实施方式中，由云端设备执行的步骤对应的程序指令以及数据。

[0390] 处理器1601，用于执行前述图7a所示实施例中任一实施例所示的由语音处理设备执行的步骤。或者用于执行前述图13所示实施例中任一实施例所示的由云端设备执行的步骤。

[0391] 通信接口1603可以用于进行数据的接收和发送，用于执行前述图7a或图13所示实施例中任一实施例中与获取、发送、接收相关的步骤。

[0392] 一种实现方式中，云端设备可以包括相对于图16更多或更少的部件，本申请对此仅仅是示例性说明，并不作限定。

[0393] 在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的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，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0394] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目

的。

[0395] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。

[0396] 当使用软件实现所述集成的单元时，可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(digital subscriber line，DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质，(例如，软盘、硬盘、磁带)、光介质(例如，DVD)、或者半导体介质(例如固态硬盘(solid state disk，SSD))等。

[0397] 本申请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术语在适当情况下可以互换，这仅仅是描述本申请的实施例中对相同属性的对象在描述时所采用的区分方式。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，以便包含一系列单元的过程、方法、系统、产品或设备不必限于那些单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它单元。

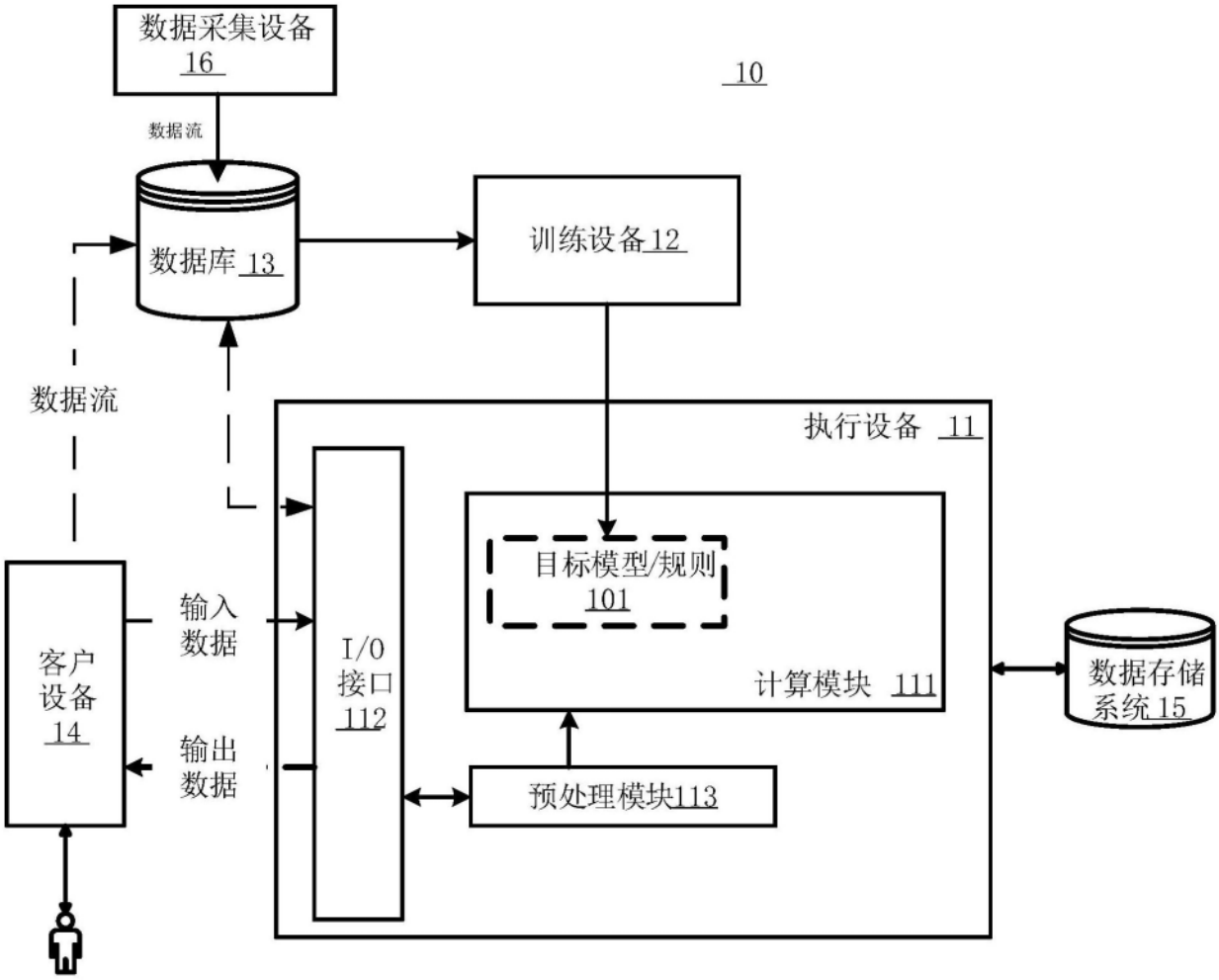

数据采集设备
16
10
数据流
数据库 13
训练设备12
数据流
执行设备 11
目标模型/规则
101
输入
数据
I/O
接口
112
客户
设备
14
输出
数据
计算模块 111
数据存储
系统 15
预处理模块113

图1

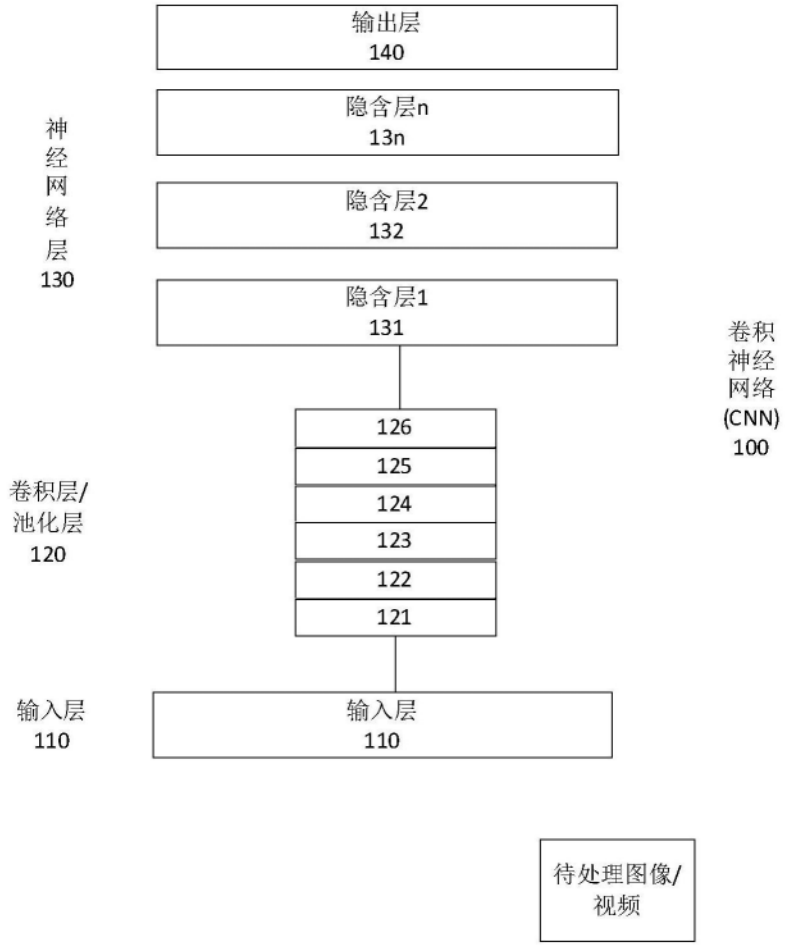

输出层
140
隐含层n
13n
隐含层2
132
隐含层1
131
神经网络层
130
卷积神经网络(CNN)
100
126
125
124
123
122
121
卷积层/池化层
120
输入层
110
输入层
110
待处理图像/视频

图2

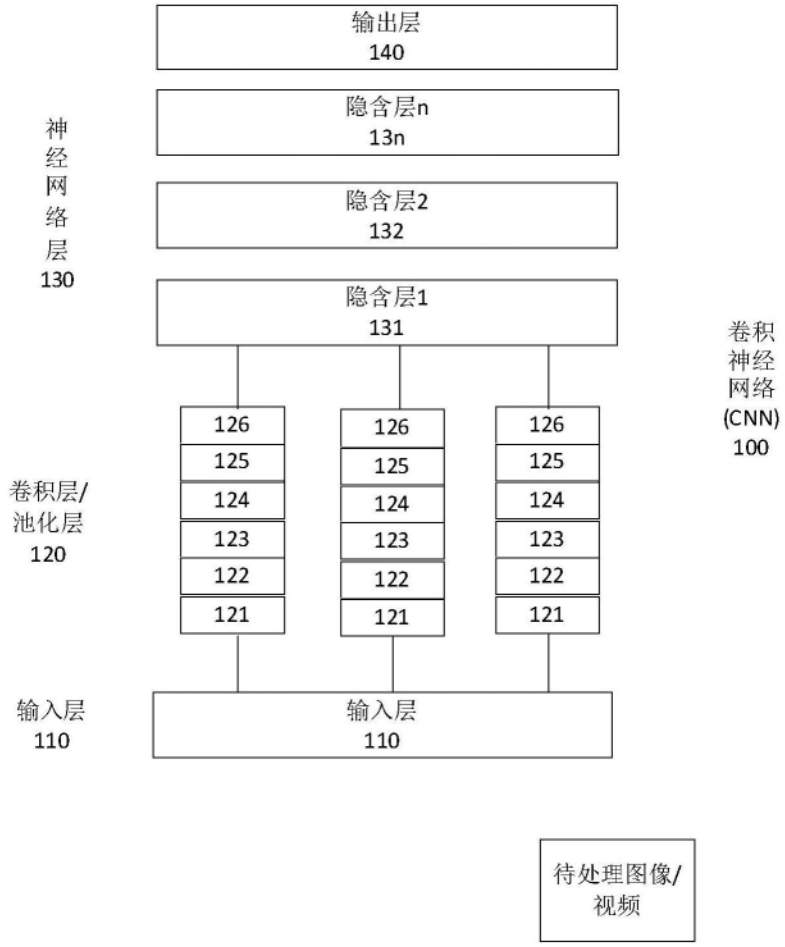

输出层
140
隐含层n
13n
隐含层2
132
隐含层1
131
神经网络层
130
卷积神经网络(CNN)
100
126
125
124
123
122
121
卷积层/池化层
120
输入层
110
输入层
110
待处理图像/视频

图3

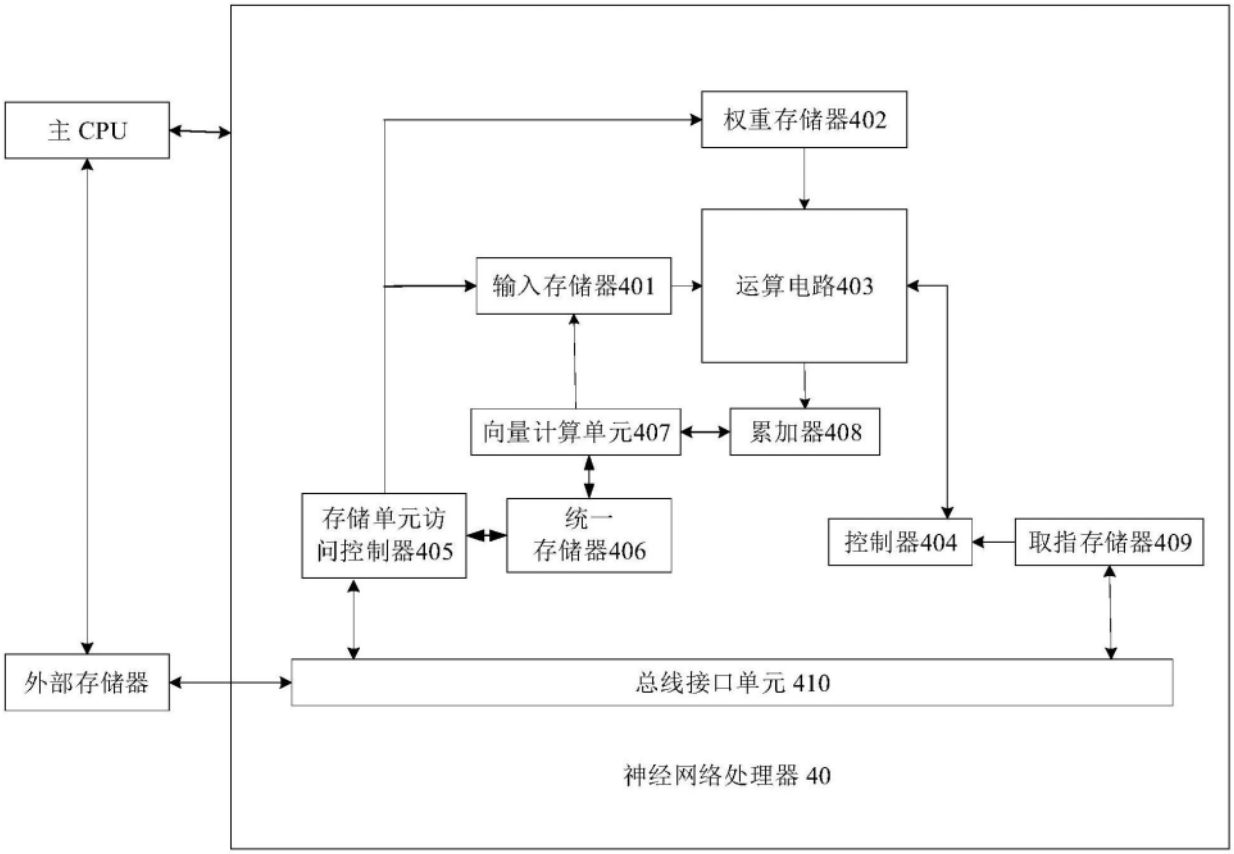

主 CPU
权重存储器402
输入存储器401
运算电路403
向量计算单元407
累加器408
存储单元访问控制器405
统一存储器406
控制器404
取指存储器409
外部存储器
总线接口单元 410
神经网络处理器 40

图4

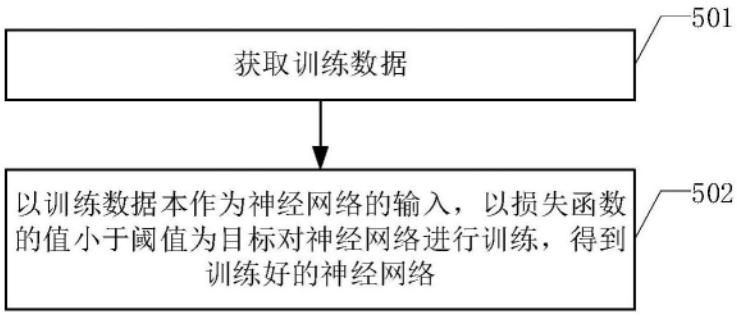

501
获取训练数据
以训练数据本作为神经网络的输入，以损失函数的值小于阈值为目标对神经网络进行训练，得到训练好的神经网络
502

图5

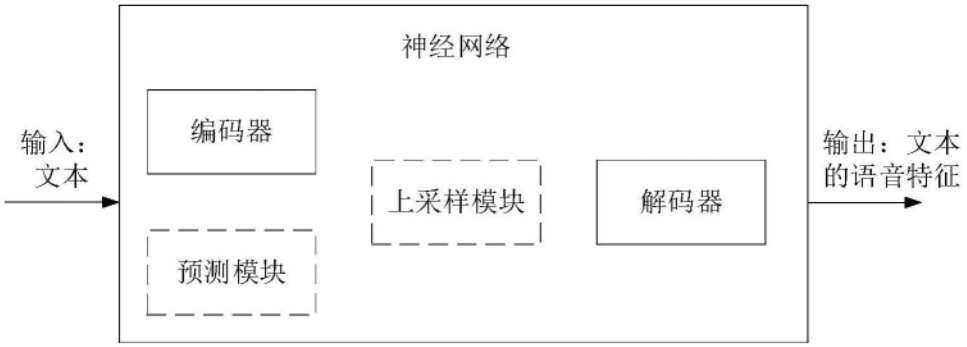

神经网络
输入：
文本
编码器
预测模块
上采样模块
解码器
输出：文本
的语音特征

图6

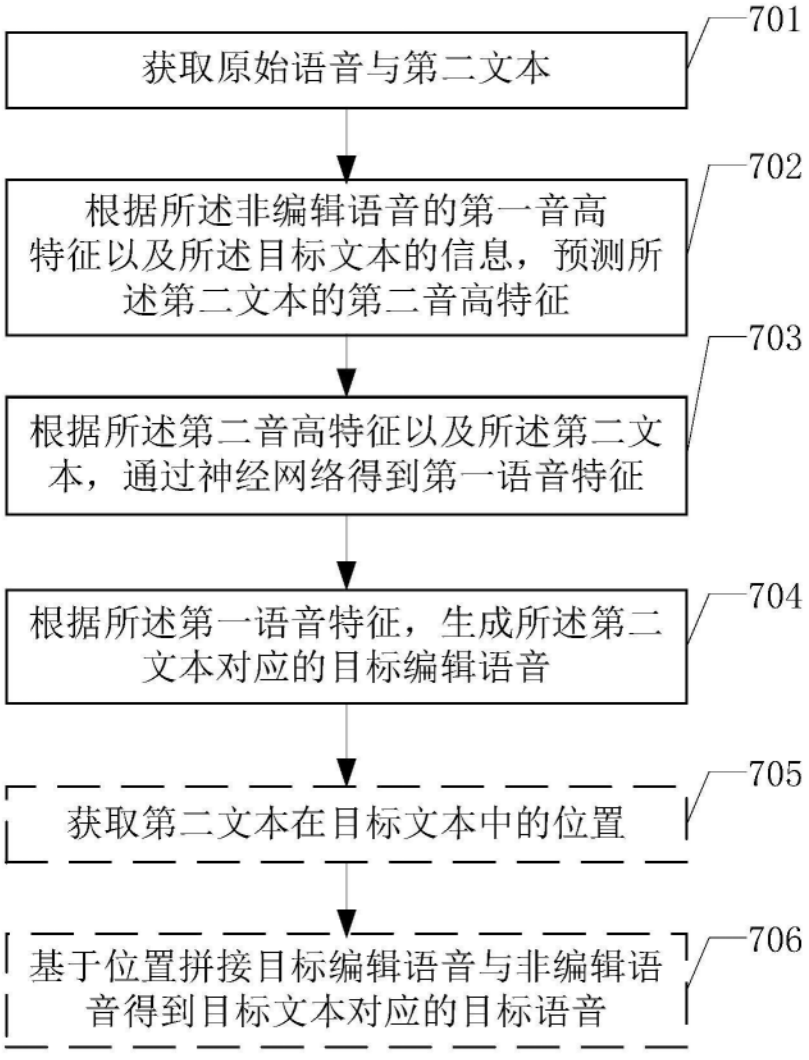

701
获取原始语音与第二文本
702
根据所述非编辑语音的第一音高特征以及所述目标文本的信息，预测所述第二文本的第二音高特征
703
根据所述第二音高特征以及所述第二文本，通过神经网络得到第一语音特征
704
根据所述第一语音特征，生成所述第二文本对应的目标编辑语音
705
获取第二文本在目标文本中的位置
706
基于位置拼接目标编辑语音与非编辑语音得到目标文本对应的目标语音

图7a

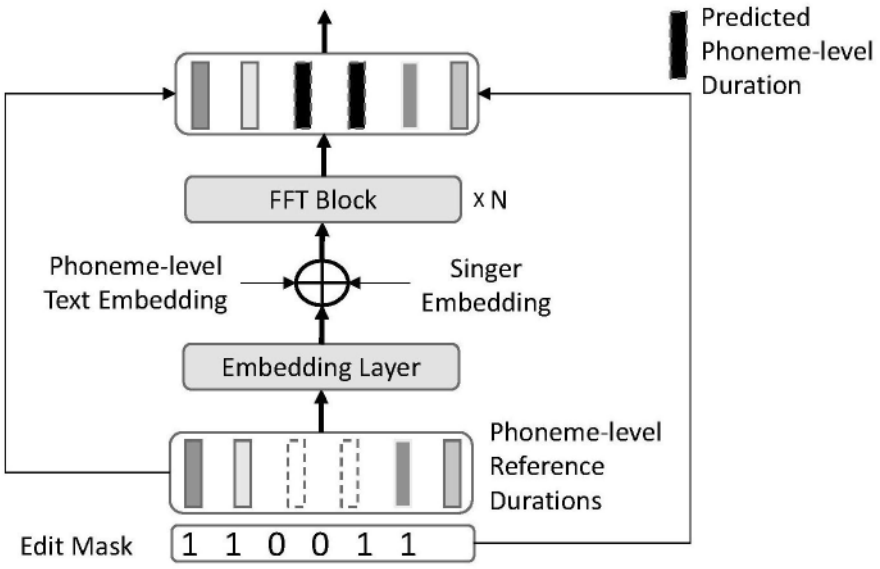

Predicted
Phoneme-level
Duration
FFT Block
x N
Phoneme-level
Text Embedding
Singer
Embedding
Embedding Layer
Phoneme-level
Reference
Durations
Edit Mask
1 1 0 0 1 1

图7b

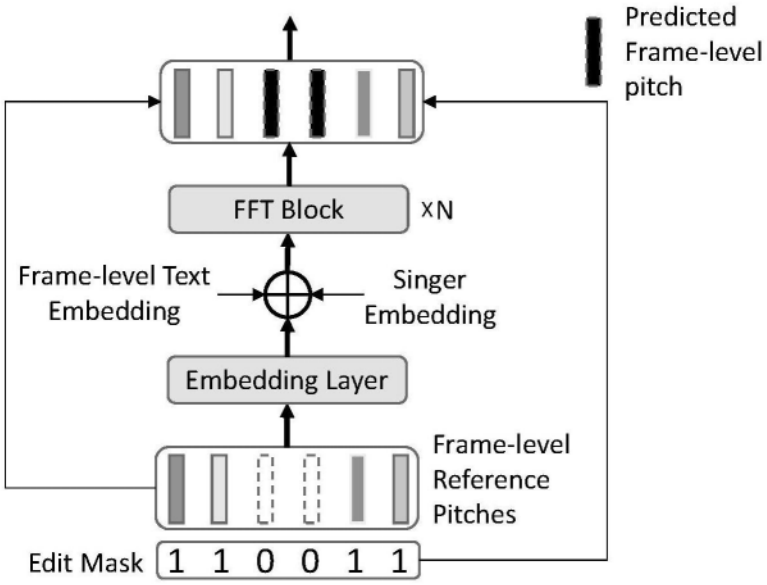

Predicted
Frame-level
pitch
FFT Block
xN
Frame-level Text
Embedding
Singer
Embedding
Embedding Layer
Frame-level
Reference
Pitches
Edit Mask
1 1 0 0 1 1

图7c

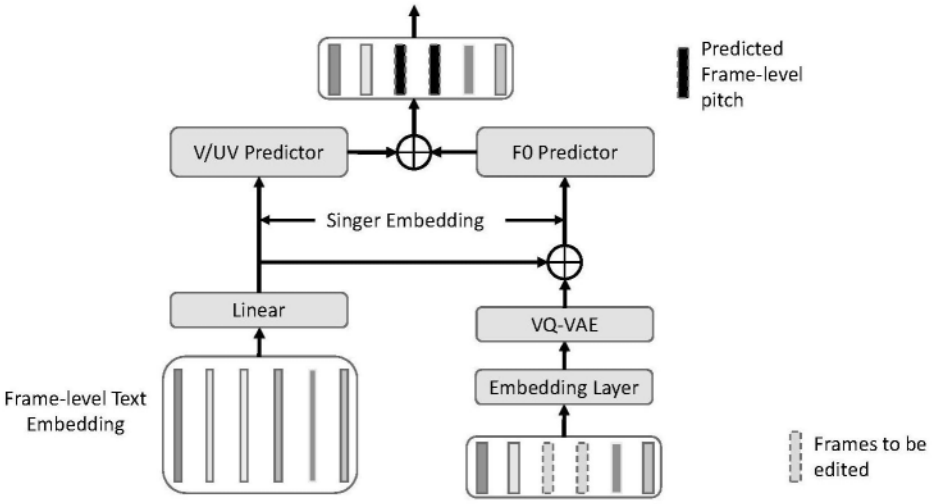

Predicted
Frame-level
pitch
V/UV Predictor
F0 Predictor
Singer Embedding
Linear
VQ-VAE
Embedding Layer
Frame-level Text
Embedding
Frames to be
edited

图7d

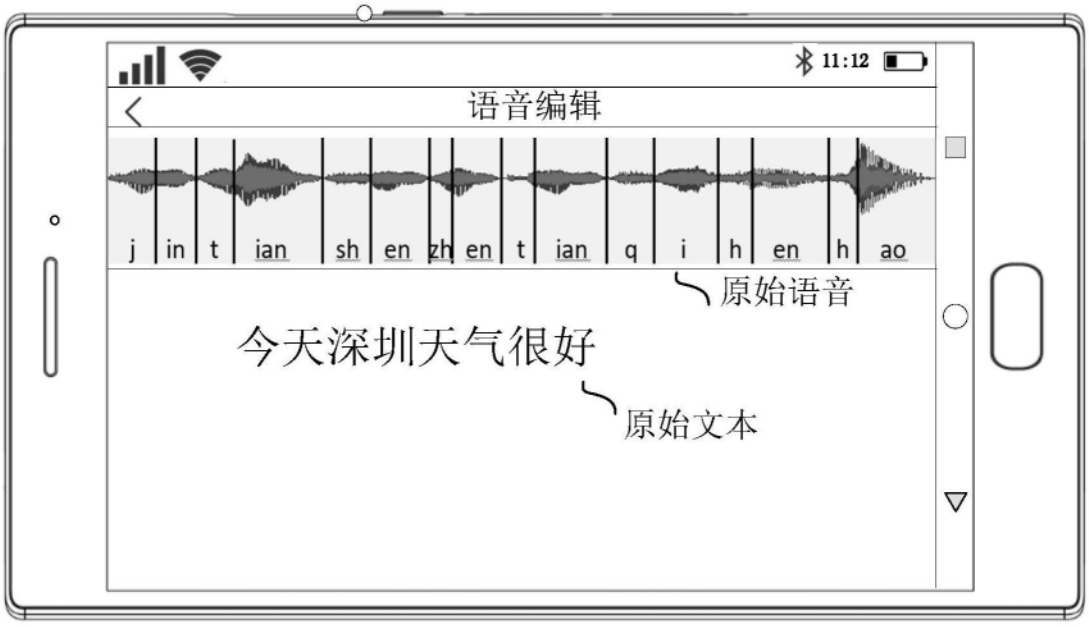

11:12
语音编辑
j in t ian sh en zh en t ian q i h en h ao
原始语音
今天深圳天气很好
原始文本

图8

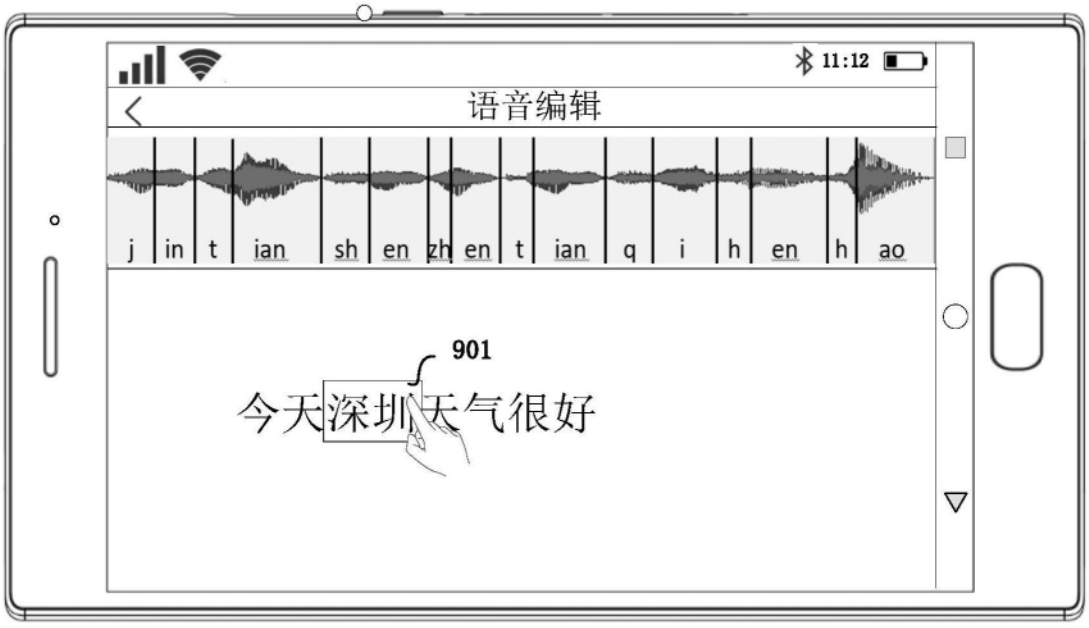

11:12
语音编辑
j in t ian sh en zh en t ian q i h en h ao
901
今天深圳天气很好

图9

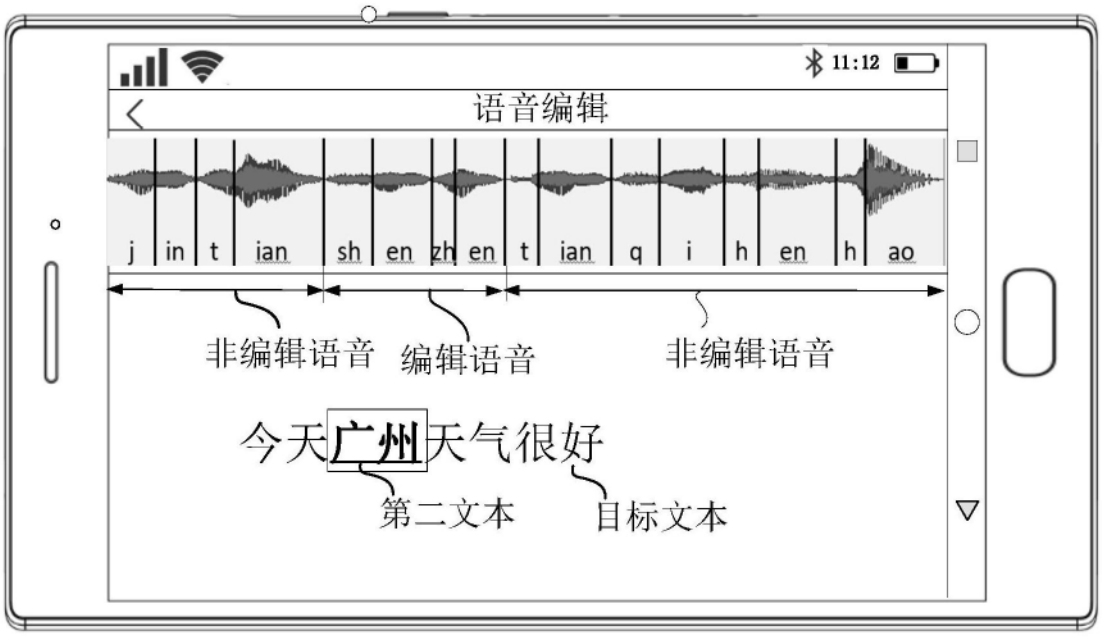

11:12
语音编辑
j in t ian sh en zh en t ian q i h en h ao
非编辑语音
编辑语音
非编辑语音
今天广州天气很好
第二文本
目标文本

图10

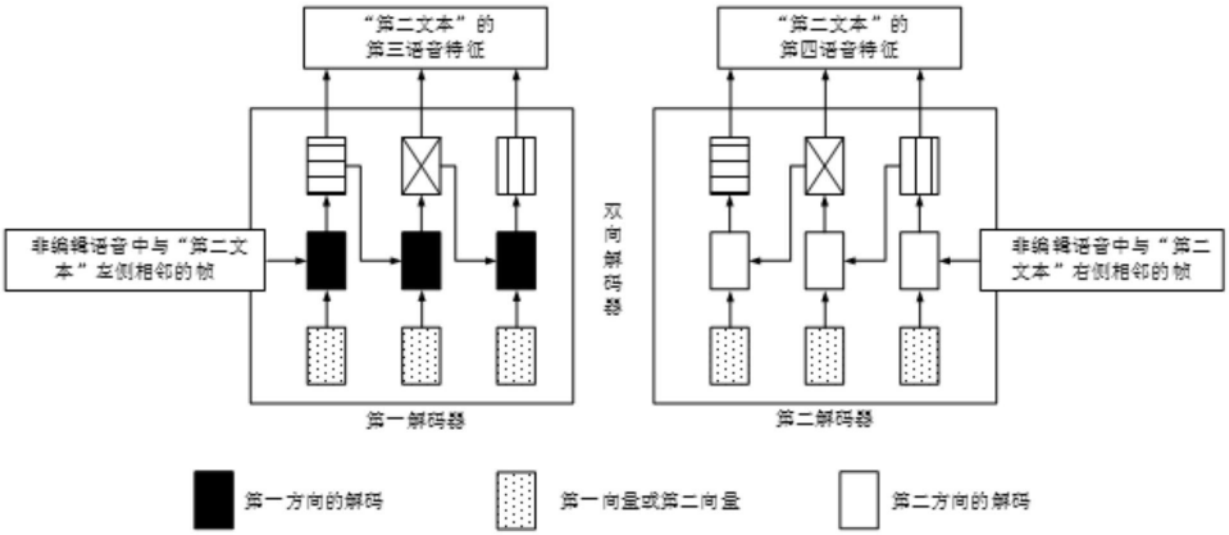

“第二文本”的
第三语音特征
“第二文本”的
第四语音特征
非编辑语音中与“第二文本”左侧相邻的帧
双向解码器
非编辑语音中与“第二文本”右侧相邻的帧
第一解码器
第二解码器
第一方向的解码
第一向量或第二向量
第二方向的解码

图11

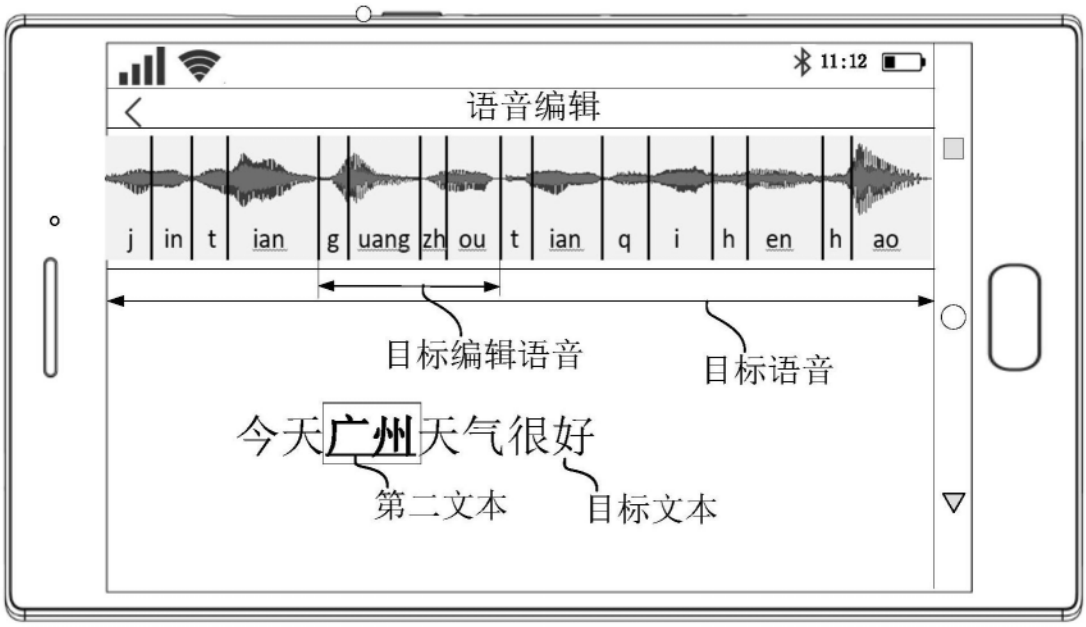

11:12
语音编辑
j in t ian g uang zh ou t ian q i h en h ao
目标编辑语音
目标语音
今天广州天气很好
第二文本
目标文本

图12

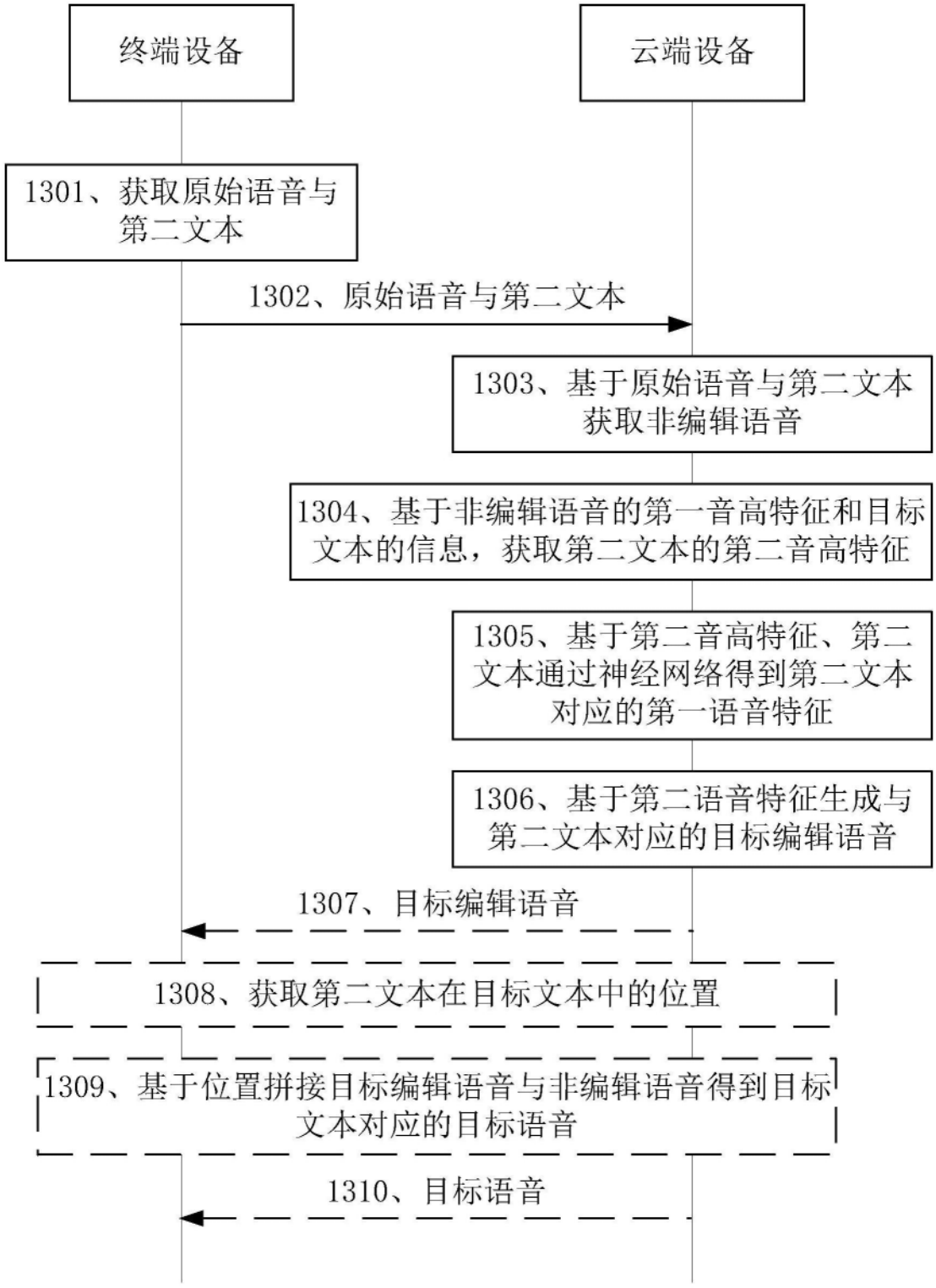

终端设备
云端设备
1301、获取原始语音与第二文本
1302、原始语音与第二文本
1303、基于原始语音与第二文本获取非编辑语音
1304、基于非编辑语音的第一音高特征和目标文本的信息，获取第二文本的第二音高特征
1305、基于第二音高特征、第二文本通过神经网络得到第二文本对应的第一语音特征
1306、基于第二语音特征生成与第二文本对应的目标编辑语音
1307、目标编辑语音
1308、获取第二文本在目标文本中的位置
1309、基于位置拼接目标编辑语音与非编辑语音得到目标文本对应的目标语音
1310、目标语音

图13

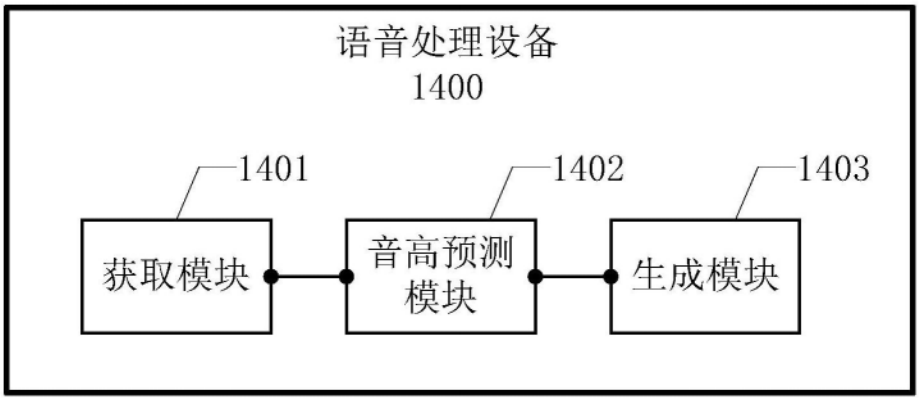

语音处理设备
1400
1401
1402
1403
获取模块
音高预测模块
生成模块

图14

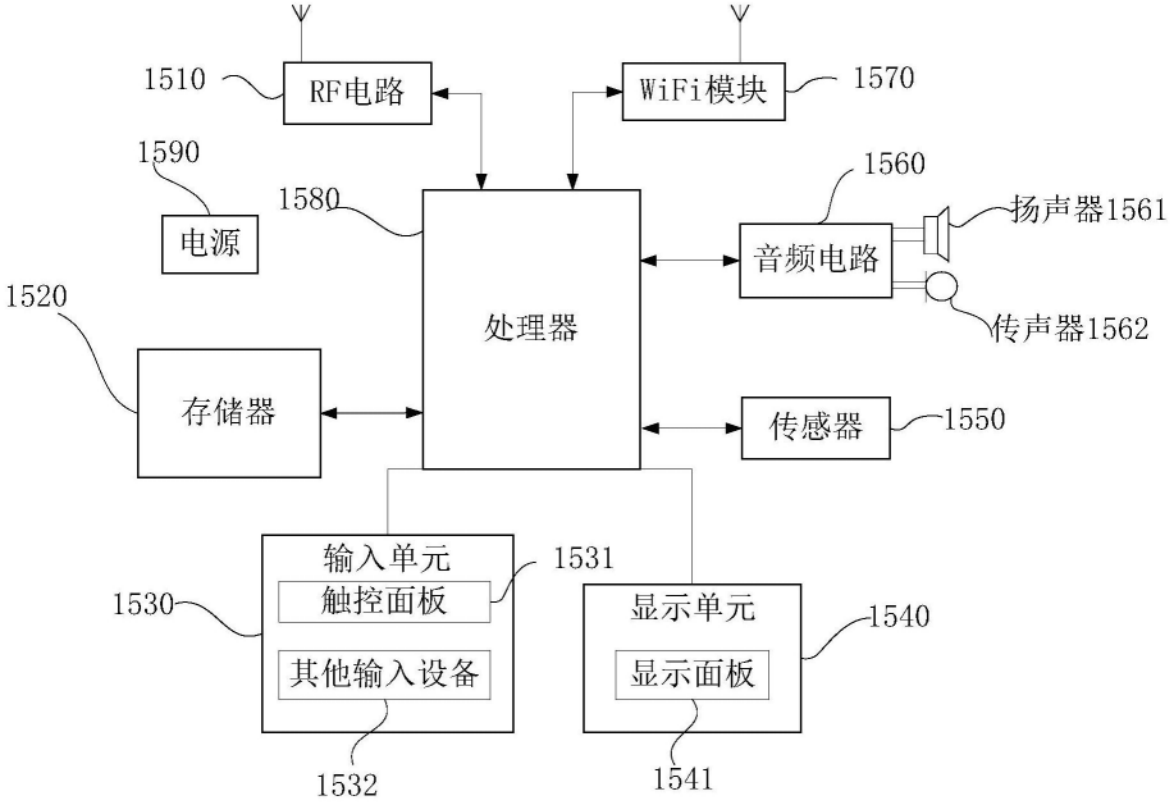

1510
RF电路
WiFi模块
1570
1590
电源
1580
处理器
1560
音频电路
扬声器1561
传声器1562
1520
存储器
传感器
1550
输入单元
1531
触控面板
1530
其他输入设备
1532
显示单元
1540
显示面板
1541

图15

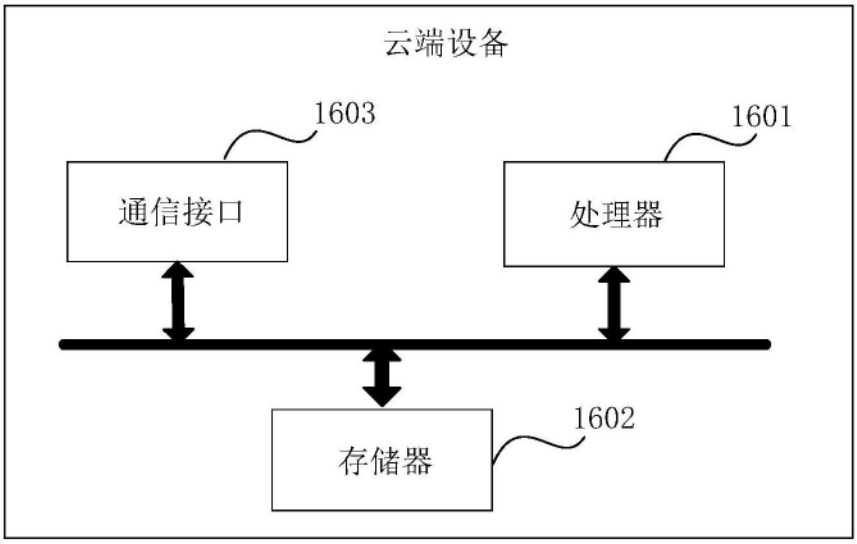



图16