

一种语音处理方法及相关设备

本申请要求于 2021 年 6 月 3 日提交中国专利局、申请号为 202110621213.6、发明名称为“一种语音处理方法及相关设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请实施例涉及人工智能领域与音频应用领域，尤其涉及一种语音处理方法及相关设备。

背景技术

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。

目前，语音编辑具有非常重要的实用意义。比如，在用户录制短视频、老师在录制授课语音等场景下，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或老师方便又快速地修正原始语音中的错误内容，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。

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听感较差。

发明内容

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。

本申请实施例第一方面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，该方法包括：获取原始语音与第二文本，第二文本为目标文本中除了第一文本以外的文本，目标文本与原始语音对应的原始文本都包括第一文本，第一文本在原始语音中对应的语音为非编辑语音；基于非编辑语音获取第一语音特征；基于第一语音特征与第二文本通过神经网络得到第二文本对应的第二语音特征；基于第二语音特征生成第二文本对应的目标编辑语音。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。

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。

本申请实施例中，通过第一文本在原始语音中的第一语音特征获取目标文本中第二文本对应的第二语音特征，即通过参考原始文本中第一文本的第一语音特征生成目标文本中第二文本的第二语音特征，进而实现目标编辑语音的听感与原始语音的听感类似，提升用户体验。

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。

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。

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。

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。

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：接收来自用户的编辑请求，编辑请求中包括原始语音与第二文本。或者编辑请求中包括原始语音与目标文本。该目标文本可以理解为是用户想要生成语音对应的文本。

该种可能的实现方式中，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。

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：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。

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。

可选地，在第一方面的一种可能的实现方式中，上述步骤：基于非编辑语音获取第一语音特征，包括：获取非编辑语音中的至少一个语音帧；基于至少一个语音帧获取第一语音特征，第一语音特征用于表示至少一个语音帧的特征，第一语音特征为特征向量或序列。另外，还可以获取目标语音（方式与前述类似），为了保证非编辑语音与目标编辑语音的衔接处更加平缓，在多个语音帧的情况下，选取的语音帧对应的文本可以与第二文本相近。

该种可能的实现方式中，一方面，通过非编辑语音中的语音帧获取第一语音特征，可以使得生成的目标编辑语音具有与非编辑语音相同或相近的语音特征，减少原始语音与目标编辑语音的听感差异。另一方面，在多个语音帧的情况下，选取的语音帧对应的文本可以与第二文本相近，进而在生成目标语音时，使得目标编辑语音与非编辑语音的衔接处更加平缓。另外，还可以通过非物理量的方式，例如，序列、向量的方式等体现语音特征。

可选地，在第一方面的一种可能的实现方式中，上述步骤中的至少一个语音帧对应的文本为第一文本中与第二文本相邻的文本。即第一语音特征对应的非编辑语音在目标语音中与非编辑语音相邻。

该种可能的实现方式中，通过第二文本的上下文的第一语音特征生成第二文本的语音特征，使得第二语音特征更加结合了上下文的第一语音特征。即通过上下文对应的语音帧预测第二文本对应的语音，可以使得第二文本的语音帧与上下文的语音帧的语音特征近似，实现第二文本的目标编辑语音与原始语音的听感类似。

可选地，在第一方面的一种可能的实现方式中，上述步骤：基于第一语音特征与第二文本通过神经网络得到第二文本对应的第二语音特征，包括：基于第一语音特征、目标文本以及标记信息通过神经网络得到第二文本对应的第二语音特征，标记信息用于标记目标文本中的第二文本。该标记信息也可以理解为是位置信息，用于指示第二文本在目标文本中的位置。

该种可能的实现方式中，通过引入目标文本，在后续生成第二文本对应的语音特征时，可以参考整个目标文本，避免后续生成的目标编辑语音与原始语音中的非编辑语音拼接得到的目标语音没有关注目标文本。

可选地，在第一方面的一种可能的实现方式中，上述步骤：神经网络包括编码器与解码器，基于第一语音特征与第二文本通过神经网络得到第二文本对应的第二语音特征，包括：基于第二文本，通过编码器，获取第二文本对应的第一向量；基于第一向量与第一语音特征，通过解码器，获取第二语音特征。也可以理解是将第一向量与第一语音特征输入解码器得到第二语音特征。

该种可能的实现方式中，解码器以第一语音特征为条件对第一向量解码，可以使得生成的第二语音特征与第一语音特征类似，或者说生成的第二语音特征携带有第一语音特征中的类似特征（例如韵律、音色和/或信噪比等）。

可选地，在第一方面的一种可能的实现方式中，上述步骤：基于第二文本，通过编码器，获取第二文本对应的第一向量，包括：基于目标文本，通过编码器，获取第一向量。也可以理解为是将目标文本以及位置信息输入编码器器得到第一向量，位置信息用于指示第二文本在目标文本中的位置。

该种可能的实现方式，在编码器编码过程中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。

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：基于目标文本通过预测网络预测第一时长与第二时长，第一时长为第一文本在目标文本中对应的音素时长，第二时长为第二文本在目标文本中对应的音素时长；基于第一时长与第三时长修正第二时长，以得到第一修正时长，第三时长为第一文本在原始语音中的音素时长；基于第一向量与第一语音特征，通过解码器，获取第二语音特征，包括：基于第一向量、第一语音特征与第一修正时长，通过解码器，获取第二语音特征。

该种可能的实现方式中，为了保证第二文本对应的目标编辑语音的时长与非编辑语音在语速上一致，可以对目标编辑语音的时长进行修正。

10 可选地，在第一方面的一种可能的实现方式中，上述步骤：基于第一时长与第三时长修正第二时长，以得到第一修正时长，包括：计算第三时长与第一时长的比值；基于比值与第二时长获取第一修正时长。

该种可能的实现方式中，利用第三时长与第一时长的比值修正第二时长。可以提升第二文本对应的目标编辑语音的时长与非编辑语音在语速上的一致程度。

15 可选地，在第一方面的一种可能的实现方式中，上述步骤：基于第一向量、第一语音特征与第一修正时长，通过解码器，获取第二语音特征，包括：基于第一修正时长对第一向量进行上采样，以得到第二向量；基于第二向量与第一语音特征，通过解码器，获取第二语音特征。具体的，将第二向量与第一语音特征输入解码器中，得到第二语音特征。在解码器包括串联的多个编码单元时，第二向量与第一语音特征可以是输入的同一个编码单元，也可以是输入的不同编码单元等。

20 该种可能的实现方式中，通过第一修正时长对第一向量进行上采样，也可以理解为是利用第一修正时长对第一向量进行扩充得到第二向量，使得目标编辑语音的时长与非编辑语音在语速上一致。

25 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：基于第二文本通过预测网络预测第四时长，第四时长为第二文本对应所有音素的总时长；获取原始语音的语速；基于语速修正第四时长，得到第二修正时长；基于第一向量与第一语音特征，通过解码器，获取第二语音特征，包括：基于第一向量、第一语音特征与第二修正时长，通过解码器，获取第二语音特征。

30 该种可能的实现方式中，利用原始语音的音素调整第二文本对应语音帧的时长，可以提升第二文本对应的目标编辑语音的时长与非编辑语音在语速上的一致程度。

可选地，在第一方面的一种可能的实现方式中，上述步骤：基于第一向量与第一语音特征，通过解码器，获取第二语音特征，包括：基于解码器与第一语音特征从目标文本的正序或反序解码第一向量得到第二语音特征。例如，目标文本为“今天开心”，则正序为从“今”至“心”的顺序，反序为从“心”至“今”的顺序。

35 该种可能的实现方式中，编码器可以通过文本的正序或反序方向预测第二语音特征。

可选地，在第一方面的一种可能的实现方式中，上述的第二文本在目标文本的中间区域，或者说第二文本并不在目标文本的两端。基于第一向量与第一语音特征，通过解码器，获取第二语音特征，包括：基于解码器与第一语音特征从目标文本的正序解码第一向量得到第三

语音特征；基于解码器与第一语音特征从目标文本的反序解码第一向量得到第四语音特征；基于第三语音特征与第四语音特征获取第二语音特征。

该种可能的实现方式中，解码器为双向解码器，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语音特征，并根据两种语音特征得到第二语音特征，使得第二语音特征与第一文本在原始语音中的特征更加近似，提升目标编辑语音的听觉效果。

可选地，在第一方面的一种可能的实现方式中，上述的第二文本包括第三文本和第四文本，第三语音特征为第三文本对应的语音特征，第四语音特征为第四文本对应的语音特征；基于第三语音特征与第四语音特征获取第二语音特征，包括：拼接第三语音特征与第四语音特征得到第二语音特征。

该种可能的实现方式中，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。

可选地，在第一方面的一种可能的实现方式中，上述步骤的第三语音特征为解码器基于正序得到的第二文本对应的语音特征，第四语音特征为解码器基于反序得到的第二文本对应的语音特征；基于第三语音特征与第四语音特征获取第二语音特征，包括：确定第三语音特征与第四语音特征中相似度大于第一阈值的语音特征为过渡语音特征；拼接第五语音特征与第六语音特征得到第二语音特征，第五语音特征为基于过渡语音特征在第三语音特征中的位置从第三语音特征中截取得到的，第六语音特征为基于过渡语音特征在第四语音特征中的位置从第四语音特征中截取得到的。

该种可能的实现方式中，通过两个完整语音特征中的过渡语音特征，从两个完整的语音特征中互补的方式选取第二语音特征，使得第二语音特征既参考了正序又参考了反序，提升第二语音特征与第一语音特征的相似程度。

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。

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。

可选地，在第一方面的一种可能的实现方式中，上述步骤：第一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。

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，使得后续生成的第二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。

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。

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。

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。

该种可能的实现方式中，可以根据用户的操作确定第一文本，一方面可以提升用户语音编辑的可操作性，另一方面，相较于使用重叠文本，可以参考更多非编辑语音的语音特征，提升目标编辑语音的听感。

可选地，在第一方面的一种可能的实现方式中，上述的神经网络是通过以训练数据作为神经网络的输入，以损失函数的值小于第二阈值为目标对神经网络进行训练得到，训练数据包括训练语音以及与训练语音对应的训练文本；损失函数用于指示神经网络输出的语音特征与实际语音特征之间的差异，实际语音特征由训练语音获取。

该种可能的实现方式中，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实际语音特征之间的差异。从而使得神经网络输出的第二语音特征更加准确。

可选地，在第一方面的一种可能的实现方式中，上述步骤：基于第一文本、原始文本以及原始语音确定非编辑语音，包括：确定原始文本中各个音素在原始语音的起止位置；基于起止位置与第一文本确定非编辑语音。

该种可能的实现方式中，根据音素在原始语音的起止位置与第一文本确定非编辑语音，使得确定的非编辑语音在音素维度上更加准确。

可选地，在第一方面的一种可能的实现方式中，上述的第一语音特征与第二语音特征为梅尔频谱特征。

本申请实施例第二方面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，该语音处理设备为终端设备，该方法包括：获取原始语音与第二文本，第二文本为目标文本中除了第一文本以外的文本，目标文本与原始语音对应的原始文本都包括第一文本，第一文本在原始语音中对应的语音为非编辑语音；向云端设备发送原始语音与第二文本，原始语音与第二文本用于云端设备生成第二文本对应的目标编辑语音；接收云端设备发送的目标编辑语音。

本申请实施例中，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。

可选地，在第二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：接收来自用户的编辑请求，编辑请求中包括原始语音与第二文本。或者编辑请求中包括原始语音与目标文本。该目标文本可以理解为是用户想要生成语音对应的文本。

该种可能的实现方式中，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。

本申请实施例第三方面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，该语音处理设备为云端设备，该方法包括：接

收终端设备发送的原始语音与第二文本，第二文本为目标文本中除了第一文本以外的文本，目标文本与原始语音对应的原始文本都包括第一文本，第一文本在原始语音中对应的语音为非编辑语音；基于非编辑语音获取第一语音特征；基于第一语音特征与第二文本通过神经网络得到第二文本对应的第二语音特征；基于第二语音特征生成第二文本对应的目标编辑语音。

5 本申请实施例中，通过第一文本在原始语音中的第一语音特征获取目标文本中第二文本对应的第二语音特征，即通过参考原始文本中第一文本的第一语音特征生成目标文本中第二文本的第二语音特征，进而实现目标编辑语音的听感与原始语音的听感类似，提升用户体验。

可选地，在第三方面的一种可能的实现方式中，上述步骤还包括：向终端设备发送目标编辑语音。

10 该种可能的实现方式中，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。

可选地，在第三方面的一种可能的实现方式中，上述步骤还包括：基于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。

15 该种可能的实现方式中，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。

本申请第四方面提供一种语音处理设备，该语音处理设备可以应用于用户录制短视频、老师录制授课语音等场景。其中，该语音处理设备可以是终端设备也可以是云端设备，该语音处理设备包括：获取单元，用于获取原始语音与第二文本，第二文本为目标文本中除了第一文本以外的文本，目标文本与原始语音对应的原始文本都包括第一文本，第一文本在原始语音中对应的语音为非编辑语音；获取单元，还用于基于非编辑语音获取第一语音特征；处理单元，用于基于第一语音特征与第二文本通过神经网络得到第二文本对应的第二语音特征；生成单元，用于基于第二语音特征生成第二文本对应的目标编辑语音。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。

25 可选地，在第四方面的一种可能的实现方式中，上述的获取单元，具体用于接收终端设备发送的原始语音与第二文本。语音处理设备还包括：发送单元，用于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。

30 可选地，在第四方面的一种可能的实现方式中，上述的获取单元，具体用于接收终端设备发送的原始语音与目标文本。生成单元，还用于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，语音处理设备还包括：发送单元，用于向终端设备发送目标语音。

可选地，在第四方面的一种可能的实现方式中，上述的获取单元，具体用于接收来自用户的编辑请求，编辑请求中包括原始语音与第二文本。或者编辑请求中包括原始语音与目标文本。该目标文本可以理解为用户想要生成语音对应的文本。

35 可选地，在第四方面的一种可能的实现方式中，上述的获取单元，还用于获取第二文本在目标文本中的位置；语音处理设备还包括：拼接单元，用于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。

可选地，在第四方面的一种可能的实现方式中，上述的获取单元，具体用于获取非编辑语音中的至少一个语音帧；获取单元，具体用于基于至少一个语音帧获取第一语音特征，第

一语音特征用于表示至少一个语音帧的特征，第一语音特征为特征向量或序列。

可选地，在第四方面的一种可能的实现方式中，上述的至少一个语音帧对应的文本为第一文本中与第二文本相邻的文本。

5 可选地，在第四方面的一种可能的实现方式中，上述的处理单元，具体用于基于第一语音特征、目标文本以及标记信息通过神经网络得到第二文本对应的第二语音特征，标记信息用于标记目标文本中的第二文本。

可选地，在第四方面的一种可能的实现方式中，上述的神经网络包括编码器与解码器，处理单元，具体用于基于第二文本，通过编码器，获取第二文本对应的第一向量；处理单元，具体用于基于第一向量与第一语音特征，通过解码器，获取第二语音特征。

10 可选地，在第四方面的一种可能的实现方式中，上述的处理单元，具体用于基于目标文本，通过编码器，获取第一向量。

可选地，在第四方面的一种可能的实现方式中，上述的语音处理设备还包括：第一预测单元，用于基于目标文本通过预测网络预测第一时长与第二时长，第一时长为第一文本在目标文本中对应的音素时长，第二时长为第二文本在目标文本中对应的音素时长；第一修正单元，用于基于第一时长与第三时长修正第二时长，以得到第一修正时长，第三时长为第一文本在原始语音中的音素时长；处理单元，具体用于基于第一向量、第一语音特征与第一修正时长，通过解码器，获取第二语音特征。

可选地，在第四方面的一种可能的实现方式中，上述的第一修正单元，具体用于计算第三时长与第一时长的比值；基于比值与第二时长获取第一修正时长。

20 可选地，在第四方面的一种可能的实现方式中，上述的处理单元，具体用于基于第一修正时长对第一向量进行上采样，以得到第二向量；处理单元，具体用于基于第二向量与第一语音特征，通过解码器，获取第二语音特征。具体的，处理单元，具体用于将第二向量与第一语音特征输入解码器中，得到第二语音特征。在解码器包括串联的多个编码单元时，第二向量与第一语音特征可以是输入的同一个编码单元，也可以是输入的不同编码单元等。

25 可选地，在第四方面的一种可能的实现方式中，上述的获取单元，还用于获取原始语音的语速；语音处理设备还包括：第二预测单元，用于基于第二文本通过预测网络预测第四时长，第四时长为第二文本对应所有音素的总时长；第二修正单元，用于基于语速修正第四时长，得到第二修正时长；处理单元，具体用于基于第一向量、第一语音特征与第二修正时长，通过解码器，获取第二语音特征。

30 可选地，在第四方面的一种可能的实现方式中，上述的处理单元，具体用于基于解码器与第一语音特征从目标文本的正序或反序解码第一向量得到第二语音特征。

可选地，在第四方面的一种可能的实现方式中，上述的第二文本在目标文本的中间区域，处理单元，具体用于基于解码器与第一语音特征从目标文本的正序解码第一向量得到第三语音特征；处理单元，具体用于基于解码器与第一语音特征从目标文本的反序解码第一向量得到第四语音特征；处理单元，具体用于基于第三语音特征与第四语音特征获取第二语音特征。

35 可选地，在第四方面的一种可能的实现方式中，上述的第二文本包括第三文本和第四文本，第三语音特征为第三文本对应的语音特征，第四语音特征为第四文本对应的语音特征；处理单元，具体用于拼接第三语音特征与第四语音特征得到第二语音特征。

可选地，在第四方面的一种可能的实现方式中，上述的第三语音特征为解码器基于正序得到的第二文本对应的语音特征，第四语音特征为解码器基于反序得到的第二文本对应的语音特征；处理单元，具体用于确定第三语音特征与第四语音特征中相似度大于第一阈值的语音特征为过渡语音特征；处理单元，具体用于拼接第五语音特征与第六语音特征得到第二语音特征，第五语音特征为基于过渡语音特征在第三语音特征中的位置从第三语音特征中截取得到的，第六语音特征为基于过渡语音特征在第四语音特征中的位置从第四语音特征中截取得到的。

可选地，在第四方面的一种可能的实现方式中，上述的生成单元，具体用于基于第二语音特征，通过声码器，生成目标编辑语音。

可选地，在第四方面的一种可能的实现方式中，上述的第一语音特征携带有原始语音的声纹特征。

可选地，在第四方面的一种可能的实现方式中，上述的获取单元，还用于基于目标文本、原始文本以及原始语音确定非编辑语音，获取单元，具体用于基于目标文本与原始文本确定第一文本；获取单元，具体用于基于第一文本、原始文本与原始语音确定非编辑语音。

可选地，在第四方面的一种可能的实现方式中，上述的获取单元，具体用于基于目标文本与原始文本确定重叠文本；获取单元，具体用于向用户显示重叠文本；获取单元，具体用于响应用户的第二操作，从重叠文本中确定第一文本。

可选地，在第四方面的一种可能的实现方式中，上述的神经网络是通过以训练数据作为神经网络的输入，以损失函数的值小于第二阈值为目标对神经网络进行训练得到，训练数据包括训练语音以及与训练语音对应的训练文本；损失函数用于指示神经网络输出的语音特征与实际语音特征之间的差异，实际语音特征由训练语音获取。

可选地，在第四方面的一种可能的实现方式中，上述的获取单元，具体用于确定原始文本中各个音素在原始语音的起止位置；获取单元，具体用于基于起止位置与第一文本确定非编辑语音。

可选地，在第四方面的一种可能的实现方式中，上述的第一语音特征与第二语音特征为梅尔频谱特征。

本申请第五方面提供一种语音处理设备，该语音处理设备可以应用于用户录制短视频、老师录制授课语音等场景。其中，该语音处理设备可以是终端设备。该语音处理设备包括：获取单元，用于获取原始语音与第二文本，第二文本为目标文本中除了第一文本以外的文本，目标文本与原始语音对应的原始文本都包括第一文本，第一文本在原始语音中对应的语音为非编辑语音；发送单元，用于向云端设备发送原始语音与第二文本，原始语音与第二文本用于云端设备生成第二文本对应的目标编辑语音；获取单元，还用于接收云端设备发送的目标编辑语音。其中，获取单元也可以理解为是输入单元，发送单元也可以理解为是输出单元。

可选地，在第五方面的一种可能的实现方式中，上述的获取单元，具体用于接收来自用户的编辑请求，编辑请求中包括原始语音与第二文本。或者编辑请求中包括原始语音与目标文本。该目标文本可以理解为用户想要生成语音对应的文本。

本申请第六方面提供一种语音处理设备，该语音处理设备可以应用于用户录制短视频、老师录制授课语音等场景。其中，该语音处理设备可以是云端设备，该语音处理设备包括：

接收单元，用于接收终端设备发送的原始语音与第二文本，第二文本为目标文本中除了第一文本以外的文本，目标文本与原始语音对应的原始文本都包括第一文本，第一文本在原始语音中对应的语音为非编辑语音；获取单元，用于基于非编辑语音获取第一语音特征；处理单元，用于基于第一语音特征与第二文本通过神经网络得到第二文本对应的第二语音特征；生成单元，用于基于第二语音特征生成第二文本对应的目标编辑语音。

可选地，在第六方面的一种可能的实现方式中，上述语音处理设备还包括：发送单元，用于向终端设备发送目标编辑语音。

可选地，在第六方面的一种可能的实现方式中，上述的生成单元，还用于基于目标编辑语音与非编辑语音生成目标语音；发送单元，用于向终端设备发送目标语音。

本申请第七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，或者执行前述第二方面或第二方面的任意可能的实现方式中的方法，或者执行前述第三方面或第三方面的任意可能的实现方式中的方法。

本申请第八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，或者使得该语音处理设备实现上述第二方面或第二方面的任意可能的实现方式中的方法，或者使得该语音处理设备实现上述第三方面或第三方面的任意可能的实现方式中的方法。

本申请第九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，或者使得计算机执行前述第二方面或第二方面的任意可能的实现方式中的方法，或者使得计算机执行前述第三方面或第三方面的任意可能的实现方式中的方法。

本申请第十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，或者使得计算机执行前述第二方面或第二方面的任意可能的实现方式中的方法，或者使得计算机执行前述第三方面或第三方面的任意可能的实现方式中的方法。

其中，第三、第四、第六、第七、第八、第九、第十方面或者其中任一种可能实现方式所带来的技术效果可参见第一方面或第一方面不同可能实现方式所带来的技术效果，此处不再赘述。

其中，第五、第七、第八、第九、第十方面或者其中任一种可能实现方式所带来的技术效果可参见第二方面或第二方面不同可能实现方式所带来的技术效果，此处不再赘述。

从以上技术方案可以看出，本申请实施例具有以下优点：通过第一文本在原始语音中的第一语音特征获取目标文本中第二文本对应的第二语音特征，即通过参考原始文本中第一文本的第一语音特征生成目标文本中第二文本的第二语音特征，进而实现目标编辑语音的听感与原始语音的听感类似，提升用户体验。

附图说明

图1为本申请提供的一种系统架构的结构示意图；

图2为本申请提供的一种卷积神经网络结构示意图；

图 3 为本申请提供的另一种卷积神经网络结构示意图；

图 4 为本申请提供的一种芯片硬件结构示意图；

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；

图 6 为本申请提供的一种神经网络的结构示意图；

5 图 7 为本申请提供的语音处理方法一个流程示意图；

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；

图 11 为本申请提供的一种双向解码器的结构示意图；

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；

图 13 为本申请提供的语音处理方法另一个流程示意图；

10 图 14-图 18 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

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。

15 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。

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。

20 1、神经网络

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：

$$h_{w,b}(x) = f(W^T x) = f(\sum_{s=1}^n W_s x_s + b)。$$

25 其中， $s=1、2、\dots、n$ ， 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，局部接受域可以由若干个神经单元组成的区域。

30 2、深度神经网络

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，35 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。

神经网络中的每一层的工作可以用数学表达式 $\bar{y} = \alpha(W\bar{x} + \bar{b})$ 来描述：从物理层面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 $W\bar{x}$ 完成，4 的操作由 $+\bar{b}$ 完成，5 的操作则由 $\alpha(\)$ 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。

3、卷积神经网络

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。

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。

4、循环神经网络 (RNN)

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。

5、损失函数

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), 它们是用来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。

6、从文本到语音

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。

7、声码器

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。

8、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。

9、韵律

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。

10、音素

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。

11、词向量(Embedding)

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, 例如: 用向量表示用户身份证标识号(identity document, ID)、物品 ID 等。

12、语音特征

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。

目前, 语音编辑的场景越来越多, 例如, 用户录制短视频、老师在录制授课语音等场景, 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。

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, 会导致校正后的语音前后不连贯、韵律不自然, 导致校正后的语音听感较差。

为了解决上述问题, 本申请提供一种语音编辑方法, 通过参考待修改内容的上下文对应

的第一语音特征确定修改内容的第二语音特征，并基于第二语音特征生成第二文本对应目标编辑语音，进而实现目标编辑语音的听感与原始语音的听感类似，提升用户体验。

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，
5 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畴。

首先介绍本申请实施例提供的系统架构。

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
10 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
15 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。

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
20 输入数据，所述输入数据在本申请实施例中可以包括：第一语音特征、目标文本以及标记信息，输入数据也可以包括第一语音特征与第二文本。另外，输入数据可以是用户输入的，也
25 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。

若输入数据包括第一语音特征、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。若输入数据包括第一语音特征、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
30 进行预处理，例如，将第二文本转化为音素等准备工作。

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。

最后，I/O 接口 112 将处理结果，如上述得到的第二语音特征返回给客户设备 14，从而
35 提供给用户。

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。

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。

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。

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。

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，另外一个任务是输出语音特征。

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。

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。

卷积层/池化层 120：

卷积层：

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。

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，
5 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。

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
10 预测。

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。

15 池化层：

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
20 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子区域的平均值或最大值。

25 神经网络层 130：

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
30 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。

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
35 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。

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。

下面介绍本申请实施例提供的一种芯片硬件结构。

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。

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。

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。

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。

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。

在一些实现种，向量计算单元 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。

统一存储器 406 用于存放输入数据以及输出数据。

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。

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。

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。

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。

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。

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。

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。

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。

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。

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。

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。

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。

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。

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 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。再不断缩小预测网络输出的训练文本中各个音素的时长与训练文本对应训练语音的实际时长之间的差异, 进而得到训练好的预测网络。

步骤 501, 获取训练数据。

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。

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。

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。

本申请实施例中获取训练数据可以通过直接录制发声对象发声的方式获取，也可以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。

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第二阈值为目标对神经网络进行训练，得到训练好的神经网络。

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。

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。

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。

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。

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。

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。

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从“中”到“没”的方向，第二方向或反序可以从“没”到“中”的方向。

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。

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。

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。

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。

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。

为了方便理解，下面对上述场景的几种举例进行描述：

一、对于替换场景。

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”。

二、对于插入场景。

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“今天上午深”的替换场景。即第一文本为“今圳天气很好”，第二文本为“今天上午深”。

三、对于删除场景。

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”。

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。

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：

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。

请参阅图 7，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 706。

步骤 701，获取原始语音与第二文本。

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。

5 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：

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。

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
10 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。

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。

15 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。

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）。

20 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。

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，
25 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。

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。
30 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。

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。

35 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

5 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。

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。

10 步骤 702，基于非编辑语音获取第一语音特征。

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通过前述的 PLP、LPC、MFCC 等方法从上述表现形式中提取的参数。

15 可选地，从非编辑语音中选取至少一个语音帧作为第一语音特征。进一步的，为了第二语音特征更加结合了上下文的第一语音特征。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。

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一语音特征。其中，编码模型可以是 CNN、RNN 等，具体此处不做限定。

20 另外，第一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，使得后续生成的第二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。

25 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。

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一语音特征。

30 示例性的，第一语音特征为梅尔频谱特征。

步骤 703，基于第一语音特征、第二文本通过神经网络得到第二文本对应的第二语音特征。

语音处理设备获取第一语音特征之后，可以基于第一语音特征、第二文本通过神经网络得到第二文本对应的第二语音特征。该神经网络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，再基于第一语音特征通过解码器对第一向量进行解码得到第二语音特征。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。

35

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。

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。即可以基于第一语音特征、目标文本、标记信息通过神经网络得到第二文本对应的第二语音特征。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，再基于第一语音特征通过解码器对第一向量进行解码得到第二语音特征。该标记信息用于标记目标文本中的第二文本。另外，为了保证第二文本对应的目标编辑语音的时长与非编辑语音在语速上一致，可以对目标编辑语音的时长进行修正。在一种可能实现的方式中，修正的具体步骤可以包括：通过预测网络预测总时长，该总时长为目标文本对应所有音素的总时长，将总时长拆分为第一时长与第二时长，第一时长为第一文本在目标文本对应的音素时长，第二时长为第二文本在目标文本对应的音素时长。再根据第一时长与第三时长修正第二时长得到第一修正时长，第三时长为第一文本在原始语音中的音素时长。在另一种可能实现的方式中，修正的具体步骤可以包括：基于第二文本通过预测网络预测第四时长，第四时长为第二文本对应所有音素的总时长；获取原始语音的语速；基于语速修正第四时长，得到第二修正时长；并基于第一向量、第一语音特征与第二修正时长，通过解码器，获取第二语音特征。类似的操作可以参考上述一种可能实现的方式中的描述，此处不再赘述。

也可以理解为是，通过第一文本在原始语音中的音素时长与预测网络预测出的第一文本在目标文本中的音素时长的差异修正第二文本在目标文本中的音素时长。

可选地，通过下述公式一计算差异系数。

$$\text{公式一: } S = \frac{\sum_{k=1}^n \frac{RP_k}{LP_k}}{n};$$

其中，n 为第一文本的音素数量， RP_k 为第 K 个音素在原始语音中的时长（即第三时长）， LP_k 为第 K 个音素在预测第二文本对应的时长（即第一时长），则第一修正时长=s*第二时长。

可选地，通过解码器获取第一向量之后，可以使用修正时长（第一修正时长或第二修正时长）对第一向量进行上采样后得到第二向量，基于第一语音特征通过解码器，解码第二向量得到第二语音特征。其中，这里的上采样可以理解为是将第一向量对应的第二时长扩展或拉伸至第二向量对应的修正时长。另外，解码器也可以通过自回归的方式获取第二语音特征，即边生成第二语音特征，边对第二语音特征进行调整。

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。

第一种，解码器是单向解码器。

解码器基于第一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二语音特征。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）。

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

- 5 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）。

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。

本申请实施例中的双向解码器有多种情况，下面分别描述：

- 10 1、双向解码器从第一方向输出的第三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。

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二语音特征。

- 15 第一解码器基于第一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三语音特征（以下称为 LR）。第二解码器基于第一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）。并根据第三语音特征与第四语音特征生成第二语音特征。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。

- 20 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 $m, m < n, \dots$ ），或者将 LR 与 RL 中差值最小的帧作为过渡帧。则第二语音特征的 N 帧可以包括 LR 中的前 m 帧与 RL 中的后 n-m 帧，或者第二语音特征的 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。

- 25 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 作为第二语音特征。

2、双向解码器从第一方向输出的第三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。

- 35 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二语音特征。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。

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 得到第二语音特征。

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语音特征，具体此处不做限定。

5 步骤 704，基于第二语音特征生成与第二文本对应的目标编辑语音。

语音处理设备获取第二语音特征之后，可以根据声码器将第二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。

10 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。

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。

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。

15 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。

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。

20 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。

25 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。

30 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。

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。

35 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 7 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。

本申请实施例中，一方面，通过第一文本在原始语音中的第一语音特征获取目标文本中第二文本对应的第二语音特征，即通过参考原始文本中第一文本的第一语音特征生成目标文本中第二文本的第二语音特征，进而实现目标编辑语音/目标语音（即编辑语音）的听感与原始语音的听感类似，提升用户体验。另一方面，通过修正目标编辑语音的时长，使得目标语音与原始语音的语速类似，提升用户体验。另一方面，可以通过直接修改原始文本的方式修改原始语音，提升用户对于语音编辑的可操作性，并且编辑后目标编辑语音同原始语音在音色、韵律等维度高度相似。另一方面，生成目标语音时，并未修改非编辑语音，且目标编辑语音的第二语音特征与非编辑语音的第一语音特征类似，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。

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。

实施例二：终端设备与云端设备共同执行语音处理方法。

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。

步骤 1301，终端设备获取原始语音与第二文本。

本实施例中终端设备执行的步骤 1301 与前述图 7 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。

步骤 1302，终端设备向云端设备发送原始语音与第二文本。

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。

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。

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。

本实施例中云端设备执行的步骤 1303 与前述图 7 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。

步骤 1304，云端设备基于非编辑语音获取第一语音特征。

步骤 1305，云端设备基于第一语音特征、第二文本通过神经网络得到第二文本对应的第二语音特征。

步骤 1306，云端设备基于第二语音特征生成与第二文本对应的目标编辑语音。

本实施例中云端设备执行的步骤 1304 至步骤 1306 与前述图 7 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。

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。

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。

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。

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。

本实施例中的步骤 1308、步骤 1309 与前述图 7 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。

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。

5 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。

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。

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。

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，且目标编辑语音的第二语音特征与非编辑语音的第一语音特征类似，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。

25 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：

获取单元 1401，用于获取原始语音与第二文本，第二文本为目标文本中除了第一文本以外的文本，目标文本与原始语音对应的原始文本都包括第一文本，第一文本在原始语音中对应的语音为非编辑语音；

30 获取单元 1401，还用于基于非编辑语音获取第一语音特征；

处理单元 1402，用于基于第一语音特征与第二文本通过神经网络得到第二文本对应的第二语音特征；

生成单元 1403，用于基于第二语音特征生成第二文本对应的目标编辑语音。

可选地，本实施例中的语音处理设备还包括：

35 拼接单元 1404，用于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。

第一预测单元 1405，用于基于目标文本通过预测网络预测第一时长与第二时长，第一时长为第一文本在目标文本中对应的音素时长，第二时长为第二文本在目标文本中对应的音素

时长；

第一修正单元 1406，用于基于第一时长与第三时长修正第二时长，以得到第一修正时长，第三时长为第一文本在原始语音中的音素时长；

第二预测单元 1407，用于基于第二文本通过预测网络预测第四时长，第四时长为第二文本对应所有音素的总时长；

第二修正单元 1408，用于基于语速修正第四时长，得到第二修正时长；

在语音处理设备为云端设备时，云端设备还可以包括发送单元 1409，用于向终端设备发送目标编辑语音或目标语音。

本实施例中，语音处理设备中各单元所执行的操作与前述图 7 至图 12 所示实施例中描述的类似，此处不再赘述。

本实施例中，一方面，处理单元 1402 通过第一文本在原始语音中的第一语音特征获取目标文本中第二文本对应的第二语音特征，即处理单元 1402 通过参考原始文本中第一文本的第一语音特征生成目标文本中第二文本的第二语音特征，进而实现生成单元 1403 生成的目标编辑语音/目标语音的听感与原始语音的听感类似，提升用户体验。另一方面，第一修正单元 1406 或者第二修正单元 1408 通过修正目标编辑语音的时长，使得目标语音与原始语音的语速类似，提升用户体验。另一方面，可以通过直接修改原始文本的方式修改原始语音，提升用户对于语音编辑的可操作性，并且编辑后目标编辑语音同原始语音在音色、韵律等维度高度相似。另一方面，生成目标语音时，并未修改非编辑语音，且目标编辑语音的第二语音特征与非编辑语音的第一语音特征类似，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。

请参阅图 15，本申请实施例中语音处理设备的另一个实施例，其中，该语音处理设备可以是终端设备。该终端设备包括：

获取单元 1501，用于获取原始语音与第二文本，第二文本为目标文本中除了第一文本以外的文本，目标文本与原始语音对应的原始文本都包括第一文本，第一文本在原始语音中对应的语音为非编辑语音；

发送单元 1502，用于向云端设备发送原始语音与第二文本，原始语音与第二文本用于云端设备生成第二文本对应的目标编辑语音；

获取单元 1501，还用于接收云端设备发送的目标编辑语音。

本实施例中，语音处理设备中各单元所执行的操作与前述图 13 所示实施例中终端设备执行步骤的描述类似，此处不再赘述。

本实施例中，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，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。

请参阅图 16，本申请实施例中语音处理设备的另一个实施例，其中，该语音处理设备可以是云端设备。该云端设备包括：

接收单元 1601，用于接收终端设备发送的原始语音与第二文本，第二文本为目标文本中除了第一文本以外的文本，目标文本与原始语音对应的原始文本都包括第一文本，第一文本

在原始语音中对应的语音为非编辑语音；

获取单元 1602，用于基于非编辑语音获取第一语音特征；

处理单元 1603，用于基于第一语音特征与第二文本通过神经网络得到第二文本对应的第二语音特征；

5 生成单元 1604，用于基于第二语音特征生成第二文本对应的目标编辑语音。

可选地，生成单元 1604，还用于基于目标编辑语音与非编辑语音生成目标语音。

可选地，本实施例中的语音处理设备还包括：

发送单元 1605，用于向终端设备发送目标编辑语音或者目标语音。

10 本实施例中，语音处理设备中各单元所执行的操作与前述图 13 所示实施例中云端设备执行步骤的描述类似，此处不再赘述。

本实施例中，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，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，生成单元 1604 生成目标语音时，并未修改非编辑语音，且目标编辑语音的第二语音特征与非编辑语音的第一语音特征类似，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。

20 请参阅图 17，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：

图 17 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 17，手机包括：射频（radio frequency, RF）电路 1710、存储器 1720、输入单元 1730、25 显示单元 1740、传感器 1750、音频电路 1760、无线保真（wireless fidelity, WiFi）模块 1770、处理器 1780、以及电源 1790 等部件。本领域技术人员可以理解，图 17 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。

下面结合图 17 对手机的各个构成部件进行具体的介绍：

30 RF 电路 1710 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 1780 处理；另外，将设计上行的数据发送给基站。通常，RF 电路 1710 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 1710 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。

存储器 1720 可用于存储软件程序以及模块，处理器 1780 通过运行存储在存储器 1720 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 1720 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 1720 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。

输入单元 1730 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 1730 可包括触控面板 1731 以及其他输入设备 1732。触控面板 1731，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 1731 上或在触控面板 1731 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 1731 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 1780，并能接收处理器 1780 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 1731。除了触控面板 1731，输入单元 1730 还可以包括其他输入设备 1732。具体地，其他输入设备 1732 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。

显示单元 1740 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 1740 可包括显示面板 1741，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 1741。进一步的，触控面板 1731 可覆盖显示面板 1741，当触控面板 1731 检测到在其上或附近的触摸操作后，传送给处理器 1780 以确定触摸事件的类型，随后处理器 1780 根据触摸事件的类型在显示面板 1741 上提供相应的视觉输出。虽然在图 17 中，触控面板 1731 与显示面板 1741 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 1731 与显示面板 1741 集成而实现手机的输入和输出功能。

手机还可包括至少一种传感器 1750，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 1741 的亮度，接近传感器可在手机移动到耳边时，关闭显示面板 1741 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。

音频电路 1760、扬声器 1761，传声器 1762 可提供用户与手机之间的音频接口。音频电路 1760 可将接收到的音频数据转换后的电信号，传输到扬声器 1761，由扬声器 1761 转换为声音信号输出；另一方面，传声器 1762 将收集的声音信号转换为电信号，由音频电路 1760 接收后转换为音频数据，再将音频数据输出处理器 1780 处理后，经 RF 电路 1710 以发送给比

如另一手机，或者将音频数据输出至存储器 1720 以便进一步处理。

WiFi 属于短距离无线传输技术，手机通过 WiFi 模块 1770 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 17 示出了 WiFi 模块 1770，但是可以理解的是，其并不属于手机的必须构成。

5 处理器 1780 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 1720 内的软件程序和/或模块，以及调用存储在存储器 1720 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 1780 可包括一个或多个处理单元；优选的，处理器 1780 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，上述调制解调处理器也可以不集成到处理器 1780 中。

10 手机还包括给各个部件供电的电源 1790（比如电池），优选的，电源可以通过电源管理系统与处理器 1780 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。

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。

15 在本申请实施例中，该终端设备所包括的处理器 1780 可以执行前述图 7 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。

参阅图 18，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 1801、存储器 1802 和通信接口 1803。该处理器 1801、存储器 1802 和通信接口 1803 通过线路互联。其中，存储器 1802 中存储有程序指令和数据。

20 存储器 1802 中存储了前述图 7 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。

处理器 1801，用于执行前述图 7 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。

25 通信接口 1803 可以用于进行数据的接收和发送，用于执行前述图 7 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。

一种实现方式中，云端设备可以包括相对于图 18 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。

30 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。

35 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。

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。

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。

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。

权 利 要 求

1. 一种语音处理方法，其特征在于，所述方法包括：

获取原始语音与第二文本，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；

基于所述非编辑语音获取第一语音特征；

基于所述第一语音特征与所述第二文本通过神经网络得到所述第二文本对应的第二语音特征；

基于所述第二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。

3. 根据权利要求 1 或 2 所述的方法，其特征在于，所述基于非编辑语音获取第一语音特征，包括：

获取所述非编辑语音中的至少一个语音帧；

基于所述至少一个语音帧获取所述第一语音特征，所述第一语音特征用于表示所述至少一个语音帧的特征，所述第一语音特征为特征向量或序列。

4. 根据权利要求 3 所述的方法，其特征在于，所述至少一个语音帧对应的文本为所述第一文本中与所述第二文本相邻的文本。

5. 根据权利要求 1 至 4 中任一项所述的方法，其特征在于，所述基于所述第一语音特征与第二文本通过神经网络得到所述第二文本对应的第二语音特征，包括：

基于所述第一语音特征、所述目标文本以及标记信息通过神经网络得到第二文本对应的第二语音特征，所述标记信息用于标记所述目标文本中的所述第二文本。

6. 根据权利要求 1 至 5 中任一项所述的方法，其特征在于，所述神经网络包括编码器与解码器，所述基于所述第一语音特征与第二文本通过神经网络得到所述第二文本对应的第二语音特征，包括：

基于所述第二文本，通过所述编码器，获取所述第二文本对应的第一向量；

基于所述第一向量与所述第一语音特征，通过所述解码器，获取所述第二语音特征。

7. 根据权利要求 6 所述的方法，其特征在于，所述基于所述第二文本，通过所述编码器，获取所述第二文本对应的第一向量，包括：

基于所述目标文本，通过所述编码器，获取所述第一向量。

8. 根据权利要求 6 或 7 所述的方法，其特征在于，所述方法还包括：

基于所述目标文本通过预测网络预测第一时长与第二时长，所述第一时长为所述第一文本在所述目标文本中对应的音素时长，所述第二时长为所述第二文本在所述目标文本中对应的音素时长；

基于所述第一时长与第三时长修正所述第二时长，以得到第一修正时长，所述第三时长为所述第一文本在所述原始语音中的音素时长；

所述基于所述第一向量与所述第一语音特征，通过所述解码器，获取所述第二语音特征，包括：

基于所述第一向量、所述第一语音特征与所述第一修正时长，通过所述解码器，获取所述第二语音特征。

5 9. 根据权利要求 6 或 7 所述的方法，其特征在于，所述方法还包括：

基于所述第二文本通过预测网络预测第四时长，所述第四时长为所述第二文本对应所有音素的总时长；

获取所述原始语音的语速；

基于所述语速修正所述第四时长，得到第二修正时长；

10 所述基于所述第一向量与所述第一语音特征，通过所述解码器，获取所述第二语音特征，包括：

基于所述第一向量、所述第一语音特征与所述第二修正时长，通过所述解码器，获取所述第二语音特征。

15 10. 根据权利要求 6 至 9 中任一项所述的方法，其特征在于，所述基于所述第一向量与所述第一语音特征，通过所述解码器，获取所述第二语音特征，包括：

基于所述解码器与所述第一语音特征从所述目标文本的正序或反序解码所述第一向量得到所述第二语音特征。

20 11. 根据权利要求 6 至 9 中任一项所述的方法，其特征在于，所述第二文本在所述目标文本的中间区域，所述基于所述第一向量与所述第一语音特征，通过所述解码器，获取所述第二语音特征，包括：

基于所述解码器与所述第一语音特征从所述目标文本的正序解码所述第一向量得到第三语音特征；

基于所述解码器与所述第一语音特征从所述目标文本的反序解码所述第一向量得到第四语音特征；

25 基于所述第三语音特征与所述第四语音特征获取所述第二语音特征。

12. 根据权利要求 11 所述的方法，其特征在于，所述第二文本包括第三文本和第四文本，所述第三语音特征为所述第三文本对应的语音特征，所述第四语音特征为所述第四文本对应的语音特征；

所述基于所述第三语音特征与所述第四语音特征获取所述第二语音特征，包括：

30 拼接所述第三语音特征与所述第四语音特征得到所述第二语音特征。

13. 根据权利要求 11 所述的方法，其特征在于，所述第三语音特征为所述解码器基于所述正序得到的所述第二文本对应的语音特征，所述第四语音特征为所述解码器基于所述反序得到的所述第二文本对应的语音特征；

所述基于所述第三语音特征与所述第四语音特征获取所述第二语音特征，包括：

35 确定所述第三语音特征与所述第四语音特征中相似度大于第一阈值的语音特征为过渡语音特征；

拼接第五语音特征与第六语音特征得到所述第二语音特征，所述第五语音特征为基于所述过渡语音特征在所述第三语音特征中的位置从所述第三语音特征中截取得到的，所述第六

语音特征为基于所述过渡语音特征在所述第四语音特征中的位置从所述第四语音特征中截取得到的。

14. 根据权利要求 1 至 13 中任一项所述的方法，其特征在于，所述基于所述第二语音特征生成所述第二文本对应的目标编辑语音，包括：

5 基于所述第二语音特征，通过声码器，生成所述目标编辑语音。

15. 根据权利要求 1 至 14 中任一项所述的方法，其特征在于，所述第一语音特征携带有所述原始语音的声纹特征。

16. 一种语音处理设备，其特征在于，所述语音处理设备包括：

10 获取单元，用于获取原始语音与第二文本，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；

所述获取单元，还用于基于所述非编辑语音获取第一语音特征；

处理单元，用于基于所述第一语音特征与所述第二文本通过神经网络得到所述第二文本对应的第二语音特征；

15 生成单元，用于基于所述第二语音特征生成所述第二文本对应的目标编辑语音。

17. 根据权利要求 16 所述的设备，其特征在于，所述获取单元，还用于获取所述第二文本在所述目标文本中的位置；

所述语音处理设备还包括：

20 拼接单元，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。

18. 根据权利要求 16 或 17 所述的设备，其特征在于，所述获取单元，具体用于获取所述非编辑语音中的至少一个语音帧；

所述获取单元，具体用于基于所述至少一个语音帧获取所述第一语音特征，所述第一语音特征用于表示所述至少一个语音帧的特征，所述第一语音特征为特征向量或序列。

25 19. 根据权利要求 18 所述的设备，其特征在于，所述至少一个语音帧对应的文本为所述第一文本中与所述第二文本相邻的文本。

20. 根据权利要求 16 至 19 中任一项所述的设备，其特征在于，所述神经网络包括编码器与解码器，所述处理单元，具体用于基于所述第二文本，通过所述编码器，获取所述第二文本对应的第一向量；

30 所述处理单元，具体用于基于所述第一向量与所述第一语音特征，通过所述解码器，获取所述第二语音特征。

21. 根据权利要求 20 所述的设备，其特征在于，所述语音处理设备还包括：

35 第一预测单元，用于基于所述目标文本通过预测网络预测第一时长与第二时长，所述第一时长为所述第一文本在所述目标文本中对应的音素时长，所述第二时长为所述第二文本在所述目标文本中对应的音素时长；

第一修正单元，用于基于所述第一时长与第三时长修正所述第二时长，以得到第一修正时长，所述第三时长为所述第一文本在所述原始语音中的音素时长；

所述处理单元，具体用于基于所述第一向量、所述第一语音特征与所述第一修正时长，

通过所述解码器，获取所述第二语音特征。

22. 根据权利要求 20 或 21 所述的设备，其特征在于，所述处理单元，具体用于基于所述解码器与所述第一语音特征从所述目标文本的正序或反序解码所述第一向量得到所述第二语音特征。

5 23. 根据权利要求 20 或 21 所述的设备，其特征在于，所述第二文本在所述目标文本的中间区域，所述处理单元，具体用于基于所述解码器与所述第一语音特征从所述目标文本的正序解码所述第一向量得到第三语音特征；

所述处理单元，具体用于基于所述解码器与所述第一语音特征从所述目标文本的反序解码所述第一向量得到第四语音特征；

10 所述处理单元，具体用于基于所述第三语音特征与所述第四语音特征获取所述第二语音特征。

24. 根据权利要求 23 所述的设备，其特征在于，所述第二文本包括第三文本和第四文本，所述第三语音特征为所述第三文本对应的语音特征，所述第四语音特征为所述第四文本对应的语音特征；

15 所述处理单元，具体用于拼接所述第三语音特征与所述第四语音特征得到所述第二语音特征。

25. 根据权利要求 23 所述的设备，其特征在于，所述第三语音特征为所述解码器基于所述正序得到的所述第二文本对应的语音特征，所述第四语音特征为所述解码器基于所述反序得到的所述第二文本对应的语音特征；

20 所述处理单元，具体用于确定所述第三语音特征与所述第四语音特征中相似度大于第一阈值的语音特征为过渡语音特征；

所述处理单元，具体用于拼接第五语音特征与第六语音特征得到所述第二语音特征，所述第五语音特征为基于所述过渡语音特征在所述第三语音特征中的位置从所述第三语音特征中截取得到的，所述第六语音特征为基于所述过渡语音特征在所述第四语音特征中的位置从
25 所述第四语音特征中截取得到的。

26. 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 15 中任一项所述的方法。

27. 根据权利要求 26 所述的设备，其特征在于，所述设备还包括：

30 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。

28. 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 15 中任一项所述的方法。

29. 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
35 述计算机执行如权利要求 1 至 15 中任一项所述的方法。