(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113436600 B
(45)授权公告日 2022.12.27

(21)申请号 202110584734.9

(22)申请日 2021.05.27

(65)同一申请的已公布的文献号
申请公布号 CN 113436600 A

(43)申请公布日 2021.09.24

(73)专利权人 北京葡萄智学科技有限公司
地址 100080 北京市海淀区北四环西路9号9层908

(72)发明人 贺宇 佟子健

(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319
专利代理师 莎日娜

(51)Int.Cl.
G10L 13/02(2013.01)
G10L 13/08(2013.01)
G10L 25/24(2013.01)
G10L 25/30(2013.01)

审查员 武金花

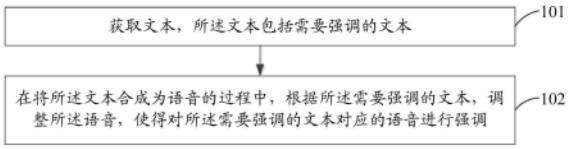
权利要求书2页 说明书11页 附图2页

(54)发明名称

一种语音合成方法及装置

(57)摘要

本申请实施例提供了一种语音合成方法及装置。所述方法包括:获取文本,所述文本包括需要强调的文本,在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音,使得对所述需要强调的文本对应的语音进行强调,使得在合成语音时,对语音进行调整,从而让文本中需要强调的部分对应的语音得到强调的效果,避免语音合成时语气平淡、没有起伏停顿,难以抓住重点的问题,实现了可控制的合成有强调的语音。



CN 113436600 B

1.一种语音合成方法,其特征在于,包括:

获取文本,所述文本包括需要强调的文本;

在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音,使得对所述需要强调的文本对应的语音进行强调;

其中,所述需要强调的文本对应的语音是通过包括前向注意力模块的语音合成模型进行强调的,在所述语音合成模型中通过将需要强调的文本对应的强调向量和当前帧的对齐函数进行点积,得到所述当前帧的强调特征;将所述当前帧的强调特征作为偏置,添加到所述前向注意力模块的输入中,得到在所述当前帧时移动到下一个音素上的概率;根据所述概率和所述文本转换成的音素序列,生成音频特征序列,基于音频特征序列进行语音强调,所述音频特征序列中,越靠近所述需要强调的文本对应的音频特征,语速越慢。

2.根据权利要求1所述的方法,其特征在于,所述在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音,使得对所述需要强调的文本对应的语音进行强调,包括:

在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音的语音属性,使得对所述需要强调的文本对应的语音进行强调,所述语音属性包括语速、基频、能量中至少一种。

3.根据权利要求2所述的方法,其特征在于,所述语音属性包括语速,所述在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音的语音属性,使得对所述需要强调的文本对应的语音进行强调,包括:

在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音的语速,使得在所述语音中,越靠近所述需要强调的文本对应的语音,语速越慢。

4.根据权利要求3所述的方法,其特征在于,所述在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音的语速,使得在所述语音中,越靠近所述需要强调的文本对应的语音,语速越慢,包括:

将所述文本输入语音合成模型得到音频特征序列,使得在所述音频特征序列中,越靠近所述需要强调的文本对应的音频特征,语速越慢;所述语音合成模型是以文本样本为输入,所述文本样本对应的音频特征序列样本为输出训练获得,所述文本样本携带需要强调的文本在所述文本样本中的位置;

基于所述音频特征序列合成所述文本对应的语音。

5.根据权利要求2所述的方法,其特征在于,所述语音属性包括语速,所述在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音的语音属性,使得对所述需要强调的文本对应的语音进行强调,包括:

在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述需要强调的文本对应的语音的语速,使得在所述语音中,所述需要强调的文本对应的语音的语速为预设语速、或逐渐增加或减少至预设语速。

6.根据权利要求2所述的方法,其特征在于,所述语音属性包括基频和/或能量,所述在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音的语音属性,使得对所述需要强调的文本对应的语音进行强调,包括:

在所述语音中,加强所述需要强调的文本对应的语音的基频和/或能量;

或者，在所述语音中，减弱所述需要强调的文本对应的语音的基频和/或能量。

7.一种语音合成装置，其特征在于，包括：

文本获取模块，用于获取文本，所述文本包括需要强调的文本；

语音合成模块，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音，使得对所述需要强调的文本对应的语音进行强调；

其中，所述需要强调的文本对应的语音是通过包括前向注意力模块的语音合成模型进行强调的，在所述语音合成模型中通过将需要强调的文本对应的强调向量和当前帧的对齐函数进行点积，得到所述当前帧的强调特征；将所述当前帧的强调特征作为偏置，添加到所述前向注意力模块的输入中，得到在所述当前帧时移动到下一个音素上的概率；根据所述概率和所述文本转换成的音素序列，生成音频特征序列，基于音频特征序列进行语音强调，所述音频特征序列中，越靠近所述需要强调的文本对应的音频特征，语速越慢。

8.根据权利要求7所述的装置，其特征在于，所述语音合成模块，包括：

属性调整子模块，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语音属性，使得对所述需要强调的文本对应的语音进行强调，所述语音属性包括语速、基频、能量中至少一种。

9.根据权利要求8所述的装置，其特征在于，所述语音属性包括语速，所述属性调整子模块，包括：

语速调整单元，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语速，使得在所述语音中，越靠近所述需要强调的文本对应的语音，语速越慢。

10.根据权利要求9所述的装置，其特征在于，所述语速调整单元，包括：

模型处理子单元，用于将所述文本输入语音合成模型得到音频特征序列，使得在所述音频特征序列中，越靠近所述需要强调的文本对应的音频特征，语速越慢；所述语音合成模型是以文本样本为输入，所述文本样本对应的音频特征序列样本为输出训练获得，所述文本样本携带需要强调的文本在所述文本样本中的位置；

语音合成子单元，用于基于所述音频特征序列合成所述文本对应的语音。

11.根据权利要求9所述的装置，其特征在于，所述语音属性包括语速，所述语速调整单元，包括：

预设语速调整子单元，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述需要强调的文本对应的语音的语速，使得在所述语音中，所述需要强调的文本对应的语音的语速为预设语速、或逐渐增加或减少至预设语速。

12.根据权利要求8所述的装置，其特征在于，所述语音属性包括基频和/或能量，所述属性调整子模块，包括：

加强单元，用于在所述语音中，加强所述需要强调的文本对应的语音的基频和/或能量；

或者，减弱单元，用于在所述语音中，减弱所述需要强调的文本对应的语音的基频和/或能量。

一种语音合成方法及装置

技术领域

[0001] 本申请涉及语音合成技术领域,特别是涉及语音合成方法、语音合成装置。

背景技术

[0002] 随着人工智能技术的高速发展,各行各业都在积极转型,将更多的业务从线下搬到线上,将产品从人力向智能化转变。而在线教育,就是近些年发展较快的行业之一。目前,主要的教学形式有直播课、录播课和AI(Artificial Intelligence,人工智能)课三种。前两种方式,内容生成效率较低,智能化程度不够,并不能符合当下用户的需求。而AI课凭借高效,智能,内容多样和个性化等特点,成为目前在线教育的主要场景。其中,语音合成在虚拟老师、智能对话、语音交互等方面,都是不可或缺的技术。

[0003] 语音合成技术,又称文语转换(Text To Speech,TTS),是一种可以将任意输入文本转换成相应语音的技术。目前在很多领域都已得到了广泛应用,如车载导航、电子书阅读、智能音箱、虚拟主播等都有涉及,但是由于这些场景特殊性,我们只需要理解字面意思,并不需要挖掘字面背后的潜在含义,因此只需要保证合成的声音足够清晰流畅即可完成所需功能。这和教育场景有很大的不同。以英语教学为例,当老师说“I have an apple”时,由于重音强调的位置不同,所要表达的意思也会有所差异,同样的一句话,强调的单词不同,所要表达的含义也千差万别。而在实际教学环境中,如果语气平淡,没有起伏停顿,可能学生并不能抓住所学重点,从而造成学习效率低下,最终导致产品体验差,造成用户流失。

[0004] 因此,传统的语音合成无法可控的合成带有强调语气的语音,并不能直接应用于在线教育领域。

发明内容

[0005] 鉴于上述问题,本申请实施例提出了一种克服上述问题的语音合成方法、语音合成装置,本申请实施例能够解决语音合成无法可控的合成带有强调语气的语音的问题。

[0006] 为了解决上述问题,本申请公开了一种语音合成方法,包括:

[0007] 获取文本,所述文本包括需要强调的文本;

[0008] 在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音,使得对所述需要强调的文本对应的语音进行强调。

[0009] 可选地,所述在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音,使得对所述需要强调的文本对应的语音进行强调,包括:

[0010] 在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音的语音属性,使得对所述需要强调的文本对应的语音进行强调,所述语音属性包括语速、基频、能量中至少一种。

[0011] 可选地,所述语音属性包括语速,所述在将所述文本合成为语音的过程中,根据所述需要强调的文本,调整所述语音的语音属性,使得对所述需要强调的文本对应的语音进行强调,包括:

[0012] 在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语速，使得在所述语音中，越靠近所述需要强调的文本对应的语音，语速越慢。

[0013] 可选地，所述在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语速，使得在所述语音中，越靠近所述需要强调的文本对应的语音，语速越慢，包括：

[0014] 将所述文本输入语音合成模型得到音频特征序列，使得在所述音频特征序列中，越靠近所述需要强调的文本对应的音频特征，语速越慢；所述语音合成模型是以文本样本为输入，所述文本样本对应的音频特征序列样本为输出训练获得，所述文本样本携带需要强调的文本在所述文本样本中的位置；

[0015] 基于所述音频特征序列合成所述文本对应的语音。

[0016] 可选地，所述语音合成模型包括前向注意力模块，所述将所述文本输入语音合成模型得到音频特征序列，使得在所述音频特征序列中，越靠近所述需要强调的文本对应的音频特征，语速越慢，包括：

[0017] 将所述文本转换成音素序列；

[0018] 将所述需要强调的文本对应的强调向量和当前帧的对齐函数进行点积，得到所述当前帧的强调特征；所述强调向量表征所述需要强调的文本对应的音素在所述音素序列中的位置；

[0019] 将所述当前帧的强调特征作为偏置，添加到所述前向注意力模块的输入中，得到在所述当前帧时移动到下一个音素上的概率；

[0020] 根据所述在所述当前帧时移动到下一个音素上的概率，以及所述音素序列，生成所述音频特征序列。

[0021] 可选地，所述语音属性包括语速，所述在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语音属性，使得对所述需要强调的文本对应的语音进行强调，包括：

[0022] 在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述需要强调的文本对应的语音的语速，使得在所述语音中，所述需要强调的文本对应的语音的语速为预设语速、或逐渐增加或减少至预设语速。

[0023] 可选地，所述语音属性包括基频和/或能量，所述在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语音属性，使得对所述需要强调的文本对应的语音进行强调，包括：

[0024] 在所述语音中，加强所述需要强调的文本对应的语音的基频和/或能量；

[0025] 或者，在所述语音中，减弱所述需要强调的文本对应的语音的基频和/或能量。

[0026] 本申请实施例还公开了一种语音合成装置，包括：

[0027] 文本获取模块，用于获取文本，所述文本包括需要强调的文本；

[0028] 语音合成模块，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音，使得对所述需要强调的文本对应的语音进行强调。

[0029] 可选地，所述语音合成模块，包括：

[0030] 属性调整子模块，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语音属性，使得对所述需要强调的文本对应的语音进行强调，所述语

音属性包括语速、基频、能量中至少一种。

[0031] 可选地，所述语音属性包括语速，所述属性调整子模块，包括：

[0032] 语速调整单元，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语速，使得在所述语音中，越靠近所述需要强调的文本对应的语音，语速越慢。

[0033] 可选地，所述语速调整单元，包括：

[0034] 模型处理子单元，用于将所述文本输入语音合成模型得到音频特征序列，使得在所述音频特征序列中，越靠近所述需要强调的文本对应的音频特征，语速越慢；所述语音合成模型是以文本样本为输入，所述文本样本对应的音频特征序列样本为输出训练获得，所述文本样本携带需要强调的文本在所述文本样本中的位置；

[0035] 语音合成子单元，用于基于所述音频特征序列合成所述文本对应的语音。

[0036] 可选地，所述语音合成模型包括前向注意力模块，所述模型处理子单元，具体用于：

[0037] 将所述文本转换成音素序列；

[0038] 将所述需要强调的文本对应的强调向量和当前帧的对齐函数进行点积，得到所述当前帧的强调特征；所述强调向量表征所述需要强调的文本对应的音素在所述音素序列中的位置；

[0039] 将所述当前帧的强调特征作为偏置，添加到所述前向注意力模块的输入中，得到在所述当前帧时移动到下一个音素上的概率；

[0040] 根据所述在所述当前帧时移动到下一个音素上的概率，以及所述音素序列，生成所述音频特征序列。

[0041] 可选地，所述语音属性包括语速，所述语速调整单元，包括：

[0042] 预设语速调整子单元，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述需要强调的文本对应的语音的语速，使得在所述语音中，所述需要强调的文本对应的语音的语速为预设语速、或逐渐增加或减少至预设语速。

[0043] 可选地，所述语音属性包括基频和/或能量，所述属性调整子模块，包括：

[0044] 加强单元，用于在所述语音中，加强所述需要强调的文本对应的语音的基频和/或能量；

[0045] 或者，减弱单元，用于在所述语音中，减弱所述需要强调的文本对应的语音的基频和/或能量。

[0046] 本申请实施例包括以下优点：

[0047] 综上所述，依据本申请实施例，通过获取文本，所述文本包括需要强调的文本，在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音，使得对所述需要强调的文本对应的语音进行强调，使得在合成语音时，对语音进行调整，从而让文本中需要强调的部分对应的语音得到强调的效果，避免语音合成时语气平淡、没有起伏停顿，难以抓住重点的问题，实现了可控制的合成有强调的语音。

附图说明

[0048] 图1示出了本申请的一种语音合成方法实施例的步骤流程图；

[0049] 图2示出了本申请的另一种语音合成方法实施例的步骤流程图；

[0050] 图3示出了基于Forward Attention的语音合成模型的架构示意图；

[0051] 图4示出了本申请的一种语音合成装置实施例的结构框图。

具体实施方式

[0052] 为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本申请作进一步详细的说明。

[0053] 参照图1，示出了本申请的一种语音合成方法实施例的步骤流程图，具体可以包括如下步骤：

[0054] 步骤101，获取文本，所述文本包括需要强调的文本。

[0055] 在本发明实施例中，文本包括需要进行语音合成的文本。文本中具有需要强调的文本，包括文本中的强调词、强调句，或者其他任意适用的部分，本发明实施例对此不做限制。例如，以英语教学为例，文本为“I have an apple”时，由于重音强调的位置不同，所要表达的意思也会有所差异([ ]内为强调词)：[I]have an apple，强调词为“I”，可能要表达是“我”有一个苹果，而不是别人；I[have]an apple，强调词为“have”，可能要教授have/has/had的区别；I have[an]apple，强调词为“an”，可能要表达冠词在元音(此处为a)前，应该用an而不是a；I have an[apple]，强调词为“apple”，可能要学习名词apple的用法。

[0056] 在本发明实施例中，需要强调的文本可以用特殊标记，来标记需要强调的文本在文本中的位置，也可以在文本中携带需要强调的文本在文本中所处的位置的位置信息，或者其他任意适用的形式，本发明实施例对此不做限制。

[0057] 步骤102，在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音，使得对所述需要强调的文本对应的语音进行强调。

[0058] 在本发明实施例中，语音，即语言的声音，是语言交际工具的声波形式。将文本合成为语音的方法可以包括多种，例如，基于拼接合成语音、统计参数语音合成方法、基于Attention(注意力)的seq2seq(sequence to sequence，序列转序列)模型、基于Forward Attention(前向注意力)的seq2seq模型，或者其他任意适用的实现方式，本发明实施例对此不做限制。

[0059] 在本发明实施例中，文本可以合成为对应的语音。为了可控制的合成带有强调的语音，在一般的文本合成语音的基础上，根据需要强调的文本，调整文本对应的语音，使得对需要强调的文本对应的语音进行强调，也就是说，在播放该语音时，使需要强调的文本对应的语音可以具有强调的效果。

[0060] 在本发明实施例中，根据所述需要强调的文本，调整所述语音，使得对所述需要强调的文本对应的语音进行强调的具体实现方式可以包括多种，例如，根据需要强调的文本，调整语音的语音属性，使得对需要强调的文本对应的语音进行强调，语音属性包括语速、基频、能量中至少一种；或者，在语音中，在需要强调的文本对应的语音处添加预设的音效；或者，在以一种音色合成语音时，将需要强调的文本对应的语音调整为另一种音色，或者其他任意适用的实现方式，本发明实施例对此不做限制。

[0061] 综上所述，依据本申请实施例，通过获取文本，所述文本包括需要强调的文本，在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音，使得对所述需

要强调的文本对应的语音进行强调，使得在合成语音时，对语音进行调整，从而让文本中需要强调的部分对应的语音得到强调的效果，避免语音合成时语气平淡、没有起伏停顿，难以抓住重点的问题，实现了可控制的合成有强调的语音。

[0062] 在本发明的一种可选实施例中，在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音，使得对所述需要强调的文本对应的语音进行强调的一种具体实现中，可以包括：在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语音属性，使得对所述需要强调的文本对应的语音进行强调，所述语音属性包括语速、基频、能量中至少一种。

[0063] 语音具有语音属性，包括但不限于语速、基频、能量等，本发明实施例对此不做限制。

[0064] 语速，其表征的是说话速度的快慢，一般用单位时间内的音节数来衡量。在文本合成为语音的过程中，对语音的语速的控制，通常仅是通过录音人本身的音色来学习模仿，并不能像真人一样，可控的强调一句话中某些重点词句。

[0065] 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频，而其他频率较高的正弦波则为泛音。

[0066] 能量，又称强度或音量，代表声音的大小，可由声音讯号的振幅来模拟，振幅越大，代表此声音波形的音量越大。

[0067] 在对语音的语音属性进行调整时，可以调整一种或多个语音属性，例如，仅调整语音的语速、不调整语音的基频和能量，或者调整语音的语速和基频，不调整语音的能量，或者调整语音的语速和能量，不调整语音的基频，或者调整语音的语速、能量和基频，或者其他任意适用的语音属性，本发明实施例对此不做限制。

[0068] 为了使得对需要强调的文本对应的语音进行强调，调整语音的语音属性时，可以调整整个语音的语音属性，或者可以调整语音中需要强调的文本对应的语音，或者可以调整部分靠近需要强调的文本对应的语音以及需要强调的文本对应的语音，或者可以调整语音中除了需要强调的文本对应的语音之外的语音的语音属性，本发明实施例对此不做限制。

[0069] 例如，调整需要强调的文本对应的语音的语速，使其语速低于其他部分的语速，或者调整整个语音的能量，使得需要强调的文本对应的语音的能量高于其他部分的能量，或者调整除了需要强调的文本对应的语音之外的语音的基频，使得需要强调的文本对应的语音的基频高于其他部分的基频，具体可以采用任意适用的实现方式，本发明实施例对此不做限制。

[0070] 在本发明实施例中，在调整语音的语速时，可以在语音中，越靠近需要强调的文本对应的语音，语速越慢，或者可以将需要强调的文本对应的语音调整为预设语速，或者可以将需要强调的文本对应的语音调整为乱序的语速，或者渐进提高或降低到预设语速等，具体可以包括任意适用的语速调整方式，本发明实施例对此不做限制。

[0071] 在本发明实施例中，在调整语音的基频或能量时，可以提高需要强调的文本对应的语音的基频或能量，或者可以减弱需要强调的文本对应的语音的基频或能量，或者可以提高除了需要强调的文本对应的语音之外的语音的基频或能量，或者可以减弱需要强调的

文本对应的语音之外的语音的基频或能量，具体可以包括任意适用的基频或能量的调整方式，本发明实施例对此不做限制。

[0072] 在本发明的一种可选实施例中，语音属性包括语速，在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语音属性，使得对所述需要强调的文本对应的语音进行强调的一种具体实现方式中，可以包括：在将所述文本合成为语音的过程中，根据需要强调的文本，调整所述语音的语速，使得在所述语音中，越靠近所述需要强调的文本对应的语音，语速越慢。

[0073] 通过对大量包含强调语气的音频进行分析，需要强调的语音一般语速较慢，而且在语音中，越靠近需要强调的语音，语速越慢。

[0074] 在本发明实施例中，在将文本合成为语音的过程中，根据需要强调的文本，对语音的语速进行调整，具体为在原本合成的语音的语速的基础上，调整后越靠近需要强调的文本对应的语音，语速越慢。

[0075] 在本发明实施例中，根据需要强调的文本对应的语音，调整语音的语速的实现方式可以包括多种，例如，先将文本转为音素序列，其中，音素是根据语音的自然属性划分出来的最小语音单位，然后按照特定的语速将音素序列转换成音频的过程中，对语速进行调整，增加部分音素在语音中出现的帧数，越靠近需要强调的文本对应的音素，在语音中出现的帧数增加的越多，需要强调的文本本身对应的音素，在语音中出现的帧数增加的最多；或者将文本输入语音合成模型得到音频特征序列，使得在音频特征序列中，越靠近需要强调的文本对应的音频特征，语速越慢；语音合成模型是以文本样本为输入，文本样本对应的音频特征序列样本为输出训练获得，文本样本携带需要强调的文本在所述文本样本中的位置，基于音频特征序列合成文本对应的语音，或者其他任意适用的实现方式，本发明实施例对此不做限制。

[0076] 在所述语音中，越靠近所述需要强调的文本对应的语音，语速越慢，使得在合成语音时，控制语音中强调部分及其前后的语速，从而让文本中的强调部分得到强调的效果，整个语音也更加自然，实现了可控制的合成有强调的语音。

[0077] 在本发明的一种可选实施例中，语音属性包括语速，在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语音属性，使得对所述需要强调的文本对应的语音进行强调的一种具体实现方式中，可以包括：在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述需要强调的文本对应的语音的语速，使得在所述语音中，所述需要强调的文本对应的语音的语速为预设语速、或逐渐增加或减少至预设语速。

[0078] 预设语速为预先设定的语速，包括匀速的语速、或者变速的语速、或者其他任意适用的语速，本发明实施例对此不做限制。为了使语速的变化不太过突兀，可以渐进的调整语速，例如，逐渐增加或逐渐减少语速，直至调整为预设语速。

[0079] 在本发明的一种可选实施例中，语音属性包括基频和/或能量，所述在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语音属性，使得对所述需要强调的文本对应的语音进行强调的一种具体实现方式中，可以包括：在所述语音中，加强所述需要强调的文本对应的语音的基频和/或能量；或者，在所述语音中，减弱所述需要强调的文本对应的语音的基频和/或能量。

[0080] 对于语音的基频或能量，可以采取加强或减弱的方式进行调整。加强基频和/或能

量，可以让倾听者明确的得知需要强调的文本对应的语音是语音中强调的部分，而减弱基频和/或能量，也可以让倾听者更加希望能够听清楚需要强调的文本对应的语音的内容，也可以作为一种强调的方式。具体可以根据实际需要，选择加强或减弱的调整方式，本发明实施例对此不做限制。例如，在语音中，加强需要强调的文本对应的语音的基频，同时减弱需要强调的文本对应的语音的能量。

[0081] 加强强调语音的基频和/或能量的方式可以包括多种，例如，在原本的基频和/或能量的基础上增加固定值的基频和/或能量，或者在原本的基频和/或能量的基础上按照比例增加基频和/或能量，或者其他任意适用的实现方式，本发明实施例对此不做限制。减弱强调语音的基频和/或能量的方式可以包括多种，例如，在原本的基频和/或能量的基础上减弱固定值的基频和/或能量，或者在原本的基频和/或能量的基础上按照比例减弱基频和/或能量，或者其他任意适用的实现方式，本发明实施例对此不做限制。

[0082] 参照图2，示出了本申请的另一种语音合成方法实施例的步骤流程图，具体可以包括如下步骤：

[0083] 步骤201，获取文本，所述文本包括需要强调的文本。

[0084] 在本发明实施例中，此步骤的具体实现方式可以参见前述实施例中的描述，此处不做赘述。

[0085] 步骤202，将所述文本输入语音合成模型得到音频特征序列，使得在所述音频特征序列中，越靠近所述需要强调的文本对应的音频特征，语速越慢；所述语音合成模型是以文本样本为输入，所述文本样本对应的音频特征序列样本为输出训练获得，所述文本样本携带需要强调的文本在所述文本样本中的位置。

[0086] 在本发明实施例中，语音合成模型可以将文本合成为对应的语音，包括但不限于基于Attention的seq2seq模型、基于Forward Attention的seq2seq模型。语音合成模型可以将输入的文本转换成对应的音频特征序列，例如，梅尔频谱(Mel Spectrogram)的帧序列。其中，音频特征是可以表征音频的特征数据，音频特征组成的序列，记为音频特征序列。

[0087] 在本发明实施例中，通常的语音合成模型是以文本样本为输入，文本样本对应的音频特征序列样本为输出训练获得的。为了控制语速，本申请实施例提出在设计语音合成模型时，增加一个与需要强调的文本对应的输入，即需要强调的文本在文本中所处的位置，那么文本样本也需要携带需要强调的文本在文本样本中的位置。在语音合成模型对语速进行控制时，在转换得到的音频特征序列中，越靠近需要强调的文本对应的音频特征，语速越慢。

[0088] 在本发明实施例中，可选地，语音合成模型包括前向注意力模块，将所述文本输入语音合成模型得到音频特征序列，使得在所述音频特征序列中，越靠近所述需要强调的文本对应的音频特征，语速越慢的具体实现方式中，可以包括：将所述文本转换成音素序列；将所述需要强调的文本对应的强调向量和当前帧的对齐函数进行点积，得到所述当前帧的强调特征；将所述当前帧的强调特征作为偏置，添加到所述前向注意力模块的输入中，得到在所述当前帧时移动到下一个音素上的概率；根据所述在所述当前帧时移动到下一个音素上的概率，以及所述音素序列，生成所述音频特征序列。

[0089] 音素序列是由音素组成的序列，例如，如果文本为“I[have]a apple”，将文本转为音素序列/AY HH AE V AX AE P AX L/。强调向量可以表征需要强调的文本对应的音素在

音素序列中的位置，例如，由于上述文本中have是强调词，则强调向量(Emphasis Embedding)为[0 1 1 1 0 0 0 0 0]，这样就可以在解码have时，有针对性的调整对应合成语音的语速了。语音合成模型中对齐函数用于计算当前帧和音素的对应关系的概率。强调向量和当前帧的对齐函数进行点积后得到的特征数据，记为当前帧的强调特征。

[0090] 如图3所示的基于Forward Attention的语音合成模型的架构示意图，语音合成模块包括编码器、解码器、前向注意力模块。输入是一句文本序列，输出是对应的梅尔频谱的帧序列。

[0091] 编码器的目标是提取输入文本的可靠序列表示(图3中的h)。编码器的输入是单词，通过发音词典转换成音素序列，每一个音素被表示成one-hot向量和嵌入成一个连续向量，然后将one-hot向量和连续向量输入三层卷积层(Conv Layers)当中，之后加上归一化层(Norm)和Relu(Rectified Linear Unit，线性整流)激活函数。最后一层卷积层的输出输入到一个双向长短期记忆网络层(Bi-LSTM，Bi-Long Short-Term Memory)当中来生成编码后的特征(h)。

[0092] 解码器是一个自动回归的循环神经网络(Recurrent Neural Network，RNN)，它将每步每帧编码的输入序列预测为梅尔频谱图，一次预测一帧。当前每一步的预测要首先通过一个包含着2个全连接层的Pre-net(初始的网络)。Pre-net作为一个信息瓶颈，在注意力学习上起着非常重要的作用，可以用来增加泛化能力和加速收敛。Pre-net的输出和注意力模块的context(上下文)向量被串联起来通过一个2层的单向LSTM层。LSTM的输出再次和注意力模块的context向量拼接在一起，然后通过一个线性转换来预测目标频谱图。最后，目标频谱帧经过一个5层卷积的“Post-net”来预测一个残差叠加到卷积前的频谱帧上，用以改善频谱重构的整个过程。并行于频谱帧的预测，解码器LSTM的输出与注意力模块的context向量拼接在一起，投影成一个标量后传递给sigmoid激活函数，来预测输出序列是否已经完成的概率。

[0093] 注意力模块可以认为是一个计算动态权重的模块。这个权重表示解码器在不同时刻t，对应输入的每个音素具有不同的权重，也就是说在解码不同帧时，我们需要的重点关注的音素是在动态变化的。注意力模块的输入是编码模块的输出隐藏特征h以及解码器中t时刻的输出的隐状态q_t，而注意力模块的输出是context向量，即是当前时刻t的动态权重。而前向注意力模块则是受到音素序列和音频特征序列单调性对齐现象的启发，使得模型更加鲁棒，收敛更快，并且可以通过TA(Transition Agent，转移代理)机制，更灵活的控制音素是否移动，而且调节DNN(Deep Neural Networks，深度学习网络)中的bias(偏置)可以控制语速快慢。

[0094] 为了控制语速，增加强调向量作为前向注意力模块的一个输入，来重点关注需要强调的文本对应的语音的语速合成。前向注意力模块是通过一个指示器$\mu_t \in (0,1)$来指示解码器在t时刻，移动到下一个音素上的概率，这个指示器由一个全连接层和Sigmoid层组成一个DNN网络，输入为在t时刻的输出$context_t$向量、解码器在t-1时刻的输出的音频特征o_{t-1}和当前音素query q_t，即：

[0095] $\mu_t \leftarrow DNN(context_t, o_{t-1}, q_t)$。

[0096] 参考$context_t$向量的生成方式，将强调向量$e=[e_1, e_2, \ldots, e_N]$和$\hat{\alpha}_t(n)$函数(对齐函数)进行点积，得到“emphasis context”向量z_t(即当前帧的强调特征)，即：

[0097] $z_t = \sum_{n=1}^{N} \hat{\alpha}_t(n) \cdot e_n$。

[0098] 然后，将z_t作为偏置加入到前向注意力模块中，即

[0099] $\mu_t \leftarrow DNN(context_t, o_{t-1}, q_t, z_t)$。

[0100] 前向注意力模块输出在当前帧时移动到下一个音素上的概率，如果z_t和e越接近，则该概率相对越小，相应的，合成的语音的语速也越慢。根据在当前帧时移动到下一个音素上的概率，以及音素序列，生成音频特征序列。因此，在音频特征序列中，越靠近需要强调的文本对应的音频特征，语速越慢。

[0101] 在一种实现方式中，对于包括前向注意力模块的语音合成模型，考虑到帧对齐特征的思想，将注意力模块输出的context特征进行扩充为`context，作为后续解码器的输入。具体的，采用concat(连接函数)。

[0102] $`context_t = Concat(context_t, z_t)$。

[0103] 这样新构造的为`context特征就包含了强调信息，在后续解码中，根据强调信息可以对强调语音的基频和/或能量进行调整。

[0104] 步骤203，基于所述音频特征序列合成所述文本对应的语音。

[0105] 在本发明实施例中，语音合成模型可以得到音频特征序列，后续只需将这个序列输入到声码器(如WaveNet、LPCNet等)即可生成时域波形样本，也就是所需要的语音，本发明对声码器部分不展开介绍。

[0106] 综上所述，依据本申请实施例，通过获取文本，所述文本包括需要强调的文本，将所述文本输入语音合成模型得到音频特征序列，使得在所述音频特征序列中，越靠近所述需要强调的文本对应的音频特征，语速越慢；所述语音合成模型是以文本样本为输入，所述文本样本对应的音频特征序列样本为输出训练获得，所述文本样本携带需要强调的文本在所述文本样本中的位置，将所述文本输入语音合成模型得到音频特征序列，使得在所述音频特征序列中，越靠近所述强调文本对应的强调音频特征，语速越慢；所述语音合成模型是以文本样本为输入，所述文本样本对应的音频特征序列样本为输出训练获得，所述文本样本携带强调文本在所述文本样本中的位置，使得在合成语音时，控制语音中强调部分及其前后的语速，从而让文本中的强调部分得到强调的效果，整个语音也更加自然，实现了可控制的合成有强调的语音。

[0107] 需要说明的是，对于方法实施例，为了简单描述，故将其都表述为一系列的运动动作组合，但是本领域技术人员应该知悉，本申请实施例并不受所描述的运动动作顺序的限制，因为依据本申请实施例，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的运动动作并不一定是本申请实施例所必须的。

[0108] 参照图4，示出了本申请的一种语音合成装置实施例的结构框图，具体可以包括：

[0109] 文本获取模块301，用于获取文本，所述文本包括需要强调的文本；

[0110] 语音合成模块302，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音，使得对所述需要强调的文本对应的语音进行强调。

[0111] 在本发明的一种可选实施例中，所述语音合成模块，包括：

[0112] 属性调整子模块，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语音属性，使得对所述需要强调的文本对应的语音进行强调，所述语

音属性包括语速、基频、能量中至少一种。

[0113] 在本发明的一种可选实施例中，所述语音属性包括语速，所述属性调整子模块，包括：

[0114] 语速调整单元，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音的语速，使得在所述语音中，越靠近所述需要强调的文本对应的语音，语速越慢。

[0115] 在本发明的一种可选实施例中，所述语速调整单元，包括：

[0116] 模型处理子单元，用于将所述文本输入语音合成模型得到音频特征序列，使得在所述音频特征序列中，越靠近所述需要强调的文本对应的音频特征，语速越慢；所述语音合成模型是以文本样本为输入，所述文本样本对应的音频特征序列样本为输出训练获得，所述文本样本携带需要强调的文本在所述文本样本中的位置；

[0117] 语音合成子单元，用于基于所述音频特征序列合成所述文本对应的语音。

[0118] 在本发明的一种可选实施例中，所述语音合成模型包括前向注意力模块，所述模型处理子单元，具体用于：

[0119] 将所述文本转换成音素序列；

[0120] 将所述需要强调的文本对应的强调向量和当前帧的对齐函数进行点积，得到所述当前帧的强调特征；所述强调向量表征所述需要强调的文本对应的音素在所述音素序列中的位置；

[0121] 将所述当前帧的强调特征作为偏置，添加到所述前向注意力模块的输入中，得到在所述当前帧时移动到下一个音素上的概率；

[0122] 根据所述在所述当前帧时移动到下一个音素上的概率，以及所述音素序列，生成所述音频特征序列。

[0123] 在本发明的一种可选实施例中，所述语音属性包括语速，所述语速调整单元，包括：

[0124] 预设语速调整子单元，用于在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述需要强调的文本对应的语音的语速，使得在所述语音中，所述需要强调的文本对应的语音的语速为预设语速、或逐渐增加或减少至预设语速。

[0125] 在本发明的一种可选实施例中，所述语音属性包括基频和/或能量，所述属性调整子模块，包括：

[0126] 加强单元，用于在所述语音中，加强所述需要强调的文本对应的语音的基频和/或能量；

[0127] 或者，减弱单元，用于在所述语音中，减弱所述需要强调的文本对应的语音的基频和/或能量。

[0128] 综上所述，依据本申请实施例，通过获取文本，所述文本包括需要强调的文本，在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音，使得对所述需要强调的文本对应的语音进行强调，使得在合成语音时，对语音进行调整，从而让文本中需要强调的部分对应的语音得到强调的效果，避免语音合成时语气平淡、没有起伏停顿，难以抓住重点的问题，实现了可控制的合成有强调的语音。

[0129] 对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关

之处参见方法实施例的部分说明即可。

[0130] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0131] 本领域内的技术人员应明白，本申请实施例的实施例可提供为方法、装置、或计算机程序产品。因此，本申请实施例可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0132] 本申请实施例是参照根据本申请实施例的方法、终端设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程终端设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理终端设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0133] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理终端设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0134] 这些计算机程序指令也可装载到计算机或其他可编程数据处理终端设备上，使得在计算机或其他可编程终端设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程终端设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0135] 尽管已描述了本申请实施例的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本申请实施例范围的所有变更和修改。

[0136] 最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者终端设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者终端设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者终端设备中还存在另外的相同要素。

[0137] 以上对本申请所提供的一种语音合成方法、一种语音合成装置，进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

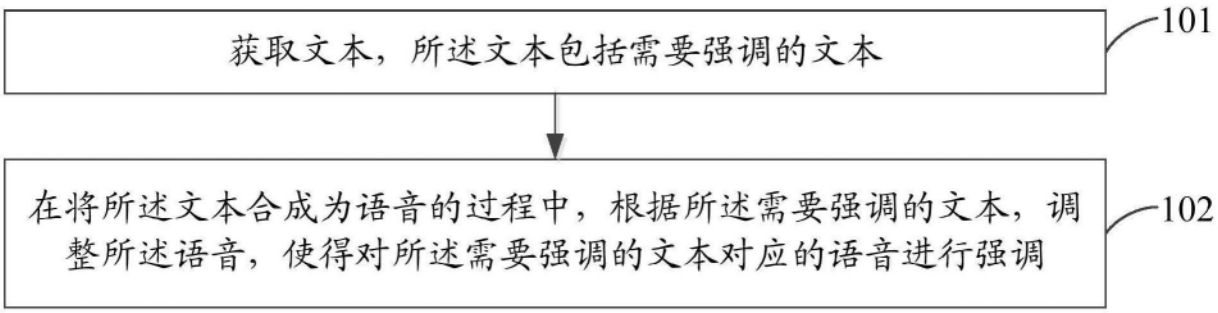

获取文本，所述文本包括需要强调的文本
101
在将所述文本合成为语音的过程中，根据所述需要强调的文本，调整所述语音，使得对所述需要强调的文本对应的语音进行强调
102

图1

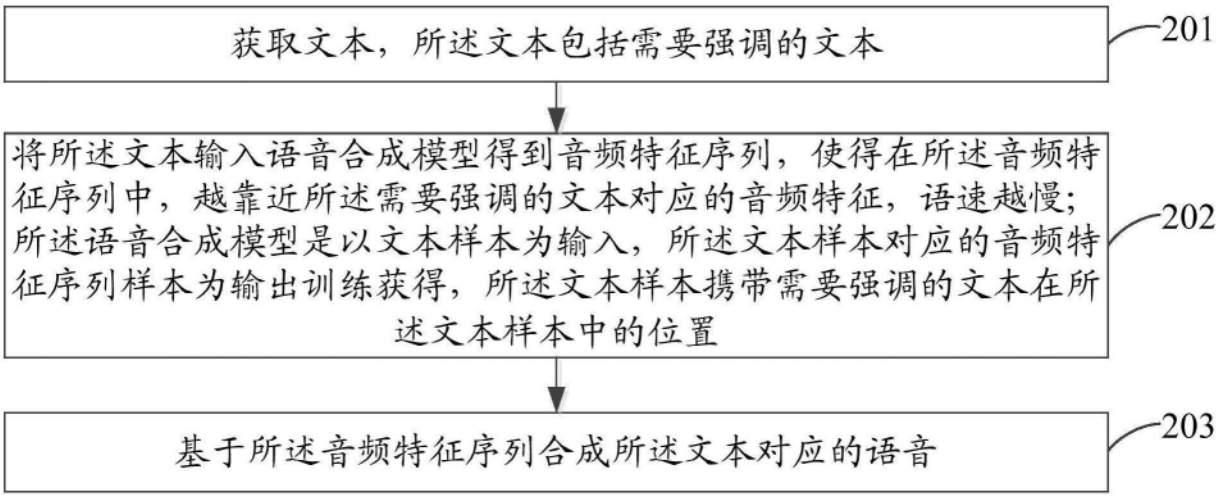

获取文本，所述文本包括需要强调的文本
201
将所述文本输入语音合成模型得到音频特征序列，使得在所述音频特征序列中，越靠近所述需要强调的文本对应的音频特征，语速越慢；所述语音合成模型是以文本样本为输入，所述文本样本对应的音频特征序列样本为输出训练获得，所述文本样本携带需要强调的文本在所述文本样本中的位置
202
基于所述音频特征序列合成所述文本对应的语音
203

图2

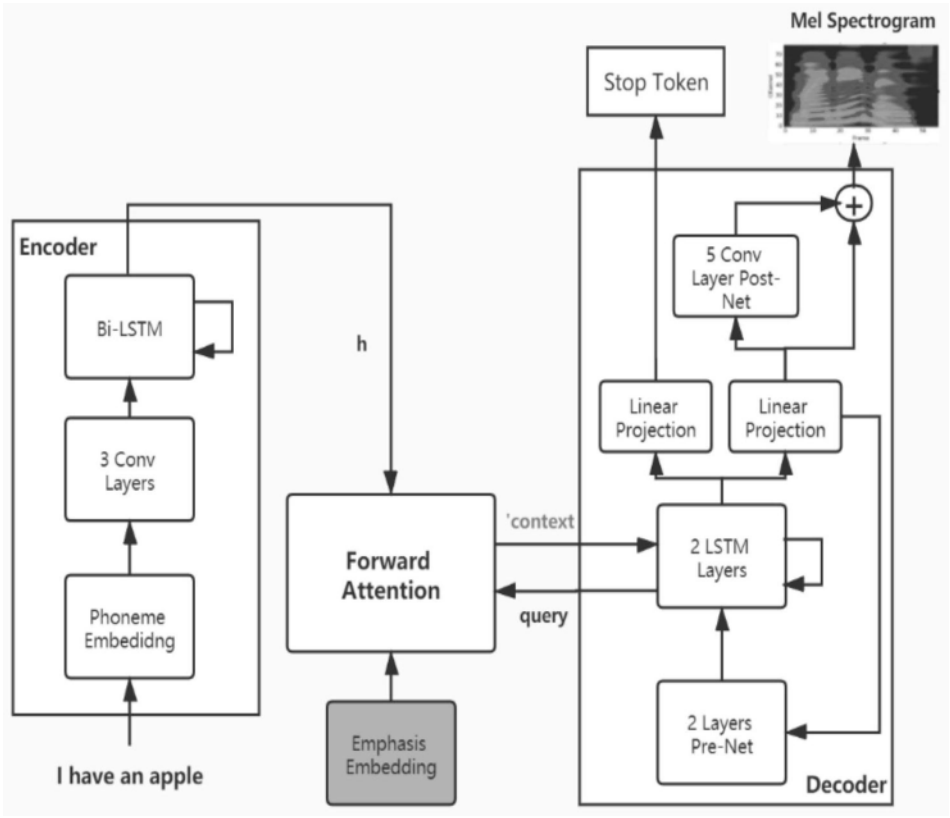

Mel Spectrogram
Stop Token
Encoder
Bi-LSTM
3 Conv Layers
Phoneme Embedidng
I have an apple
h
Forward Attention
Emphasis Embedding
'context
query
5 Conv Layer Post-Net
Linear Projection
Linear Projection
2 LSTM Layers
2 Layers Pre-Net
Decoder

图3


文本获取模块
301
语音合成模块
302

图4