

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 5 月 12 日 (12.05.2022)



(10) 国际公布号
WO 2022/095754 A1

(51) 国际专利分类号:
G10L 13/10 (2013.01)

中国北京市海淀区知春路63号中国卫星通信大厦今日头条小邮局, Beijing 100190 (CN)。

(21) 国际申请号: PCT/CN2021/126394

(22) 国际申请日: 2021 年 10 月 26 日 (26.10.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202011212351.0 2020年11月3日 (03.11.2020) CN

(71) 申请人: 北京有竹居网络技术有限公司(BEIJING YOUZHUJU NETWORK TECHNOLOGY CO. LTD.) [CN/CN]; 中国北京市平谷区林荫北街13号信息大厦802室, Beijing 101299 (CN)。

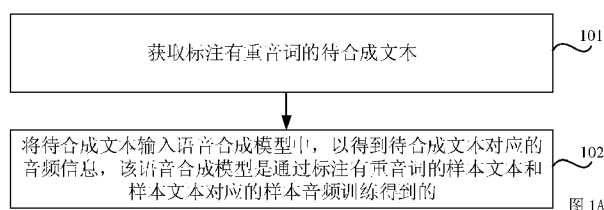
(72) 发明人: 徐晨畅(XU, Chenchang); 中国北京市海淀区知春路63号中国卫星通信大厦今日头条小邮局, Beijing 100190 (CN)。 潘俊杰(PAN, Junjie);

(74) 代理人: 中国贸促会专利商标事务所有限公司(CCPIT PATENT AND TRADEMARK LAW OFFICE); 中国北京市复兴门内大街158号远洋大厦F10层, Beijing 100031 (CN)。

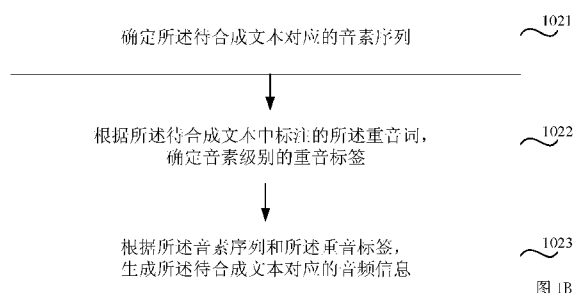
(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) Title: SPEECH SYNTHESIS METHOD AND APPARATUS, STORAGE MEDIUM, AND ELECTRONIC DEVICE

(54) 发明名称: 语音合成方法、装置、存储介质及电子设备



101 Acquire text to be synthesized labeled with accented words
102 Input the text to be synthesized into a speech synthesis model to obtain audio information corresponding to the text to be synthesized, the speech synthesis model being obtained by performing training by means of sample text labeled with accented words and sample audios corresponding to the sample text



1021 Determine a phoneme sequence corresponding to the text to be synthesized
1022 Determine, according to the accented words labeled in the text to be synthesized, accented labels of the phoneme level
1023 Generate, according to the phoneme sequence and the accented labels, audio information corresponding to the text to be synthesized

(57) Abstract: A speech synthesis method and apparatus, a storage medium, and an electronic device. The method comprises: acquiring text to be synthesized labeled with accented words (101); and inputting the text to be synthesized into a speech synthesis model to obtain audio information corresponding to the text to be synthesized, the speech synthesis model being obtained by performing training by means of sample text labeled with accented words and sample audios corresponding to the sample text (102), the speech synthesis model being used for processing the text to be synthesized in the following manner: determining a phoneme sequence corresponding to the text to be synthesized (1021); determining, according to the accented words labeled in the text to be synthesized, accented labels of the phoneme level (1022); and generating, according to the phoneme sequence and the accented labels, audio information corresponding to the text to be synthesized (1023). By the present method, a synthesized speech with accents can be obtained, and accuracy of accent pronunciation in the synthesized speech can be ensured.

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 一种语音合成方法、装置、存储介质及电子设备, 该方法包括: 获取标注有重音词的待合成文本 (101); 将待合成文本输入语音合成模型中, 以得到待合成文本对应的音频信息, 语音合成模型是通过标注有重音词的样本文本和样本文本对应的样本音频训练得到的 (102)。进一步地, 语音合成模型用于通过如下方式对待合成文本进行处理: 确定待合成文本对应的音素序列 (1021); 根据待合成文本中标注的重音词, 确定音素级别的重音标签 (1022); 根据音素序列和重音标签, 生成待合成文本对应的音频信息 (1023)。该方法能够得到带重音的合成语音, 并保证合成语音中重音发音的准确性。

语音合成方法、装置、存储介质及电子设备

相关申请的交叉引用

本申请是以申请号为 202011212351.0、申请日为 2020 年 11 月 03 日的中国申请为基础，并主张其优先权，该中国申请的公开内容在此作为整体引入本申请中。

技术领域

本公开涉及语音合成技术领域，具体地，涉及一种语音合成方法、装置、存储介质及电子设备。

背景技术

语音合成，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。传统的语音合成系统通常包括前端和后端两个模块。前端模块主要是对输入文本进行分析，提取后端模块所需要的语言学信息。后端模块根据前端分析结果，通过一定的方法生成语音波形。

但是，相关技术中的语音合成方式通常没有考虑合成语音中的重音，导致合成语音中没有重音，发音平淡，缺乏表现力。或者，相关技术中的语音合成方式通常随机选择输入文本中的词语进行重音添加，导致合成语音中的重音发音错误，无法得到较好的包含重音的语音合成结果。

发明内容

提供该发明内容部分以便以简要的形式介绍构思，这些构思将在后面的具体实施方式部分被详细描述。该发明内容部分并不旨在标识要求保护的技术方案的关键特征或必要特征，也不旨在用于限制所要求的保护的技術方案的範圍。

第一方面，本公开提供一种语音合成方法，所述方法包括：

获取标注有重音词的待合成文本；

将所述待合成文本输入语音合成模型中，以得到所述待合成文本对应的音频信息，所述语音合成模型是通过标注有重音词的样本文本和所述样本文本对应的样本音频训练得到的，所述语音合成模型用于通过如下方式对所述待合成文本进行处理：

确定所述待合成文本对应的音素序列；

根据所述待合成文本中标注的所述重音词，确定音素级别的重音标签；

根据所述音素序列和所述重音标签，生成所述待合成文本对应的音频信息。

第二方面，本公开提供一种语音合成装置，所述装置包括：

获取模块，用于获取标注有重音词的待合成文本；

- 5 合成模块，用于将所述待合成文本输入语音合成模型中，以得到所述待合成文本对应的音频信息，所述语音合成模型是通过标注有重音词的样本文本和所述样本文本对应的样本音频训练得到的，所述语音合成模型用于通过如下模块对所述待合成文本进行处理：

第一确定子模块，用于确定所述待合成文本对应的音素序列；

- 10 第二确定子模块，用于根据所述待合成文本中标注的所述重音词，确定音素级别的重音标签；

生成子模块，用于根据所述音素序列和所述重音标签，生成所述待合成文本对应的音频信息。

第三方面，本公开提供一种计算机可读介质，其上存储有计算机程序，该程序被处理装置执行时实现第一方面中所述方法的步骤。

- 15 第四方面，本公开提供一种电子设备，包括：

存储装置，其上存储有计算机程序；

处理装置，用于执行所述存储装置中的所述计算机程序，以实现第一方面中所述方法的步骤。

- 20 第五方面，本公开提供一种计算机程序产品，包括指令，所述指令在由计算机执行时使得计算机实现第一方面中所述方法的步骤。

- 25 通过上述技术方案，可以根据标注有重音词的样本文本和该样本文本对应的样本音频训练语音合成模型，训练后的语音合成模型可以根据标注有重音词的待合成文本生成包括有重音发音的音频信息。并且，由于语音合成模型是根据标注有重音词的大量样本文本训练得到的，因此相较于相关技术中随机添加重音发音的方式，一定程度上可以保证生成的音频信息的准确性。此外，语音合成模型可以在将待合成文本扩展到音素级别的情况下进行语音合成处理，因此可以使得合成语音中的重音在音素级别上可控，从而进一步提升合成语音中重音发音的准确性。

本公开的其他特征和优点将在随后的具体实施方式部分予以详细说明。

- 30 附图说明

结合附图并参考以下具体实施方式，本公开各实施例的上述和其他特征、优点及方面

将变得更加明显。贯穿附图中，相同或相似的附图标记表示相同或相似的元素。应当理解附图是示意性的，原件和元素不一定按照比例绘制。在附图中：

图 1A 和图 1B 是根据本公开一示例性实施例示出的一种语音合成方法的流程图，图 1C 是根据本公开示例性实施例的重音词确定过程的流程图，图 1D 是根据本公开示例性实
5 施例的语音合成模型确定过程的流程图；

图 2 是根据本公开一示例性实施例示出的一种语音合成方法中语音合成模型的示意图；

图 3 是根据本公开另一示例性实施例示出的一种语音合成方法中语音合成模型的示意图；

10 图 4 是根据本公开一示例性实施例示出的一种语音合成装置的框图；

图 5 是根据本公开一示例性实施例示出的一种电子设备的框图。

具体实施方式

下面将参照附图更详细地描述本公开的实施例。虽然附图中显示了本公开的某些实施
15 例，然而应当理解的是，本公开可以通过各种形式来实现，而且不应该被解释为限于这里
阐述的实施例，相反提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是，
本公开的附图及实施例仅用于示例性作用，并非用于限制本公开的保护范围。

应当理解，本公开的方法实施方式中记载的各个步骤可以按照不同的顺序执行，和/
或并行执行。此外，方法实施方式可以包括附加的步骤和/或省略执行示出的步骤。本公
20 开的范围在此方面不受限制。

本文使用的术语“包括”及其变形是开放性包括，即“包括但不限于”。术语“基于”
是“至少部分地基于”。术语“一个实施例”表示“至少一个实施例”；术语“另一实施
例”表示“至少一个另外的实施例”；术语“一些实施例”表示“至少一些实施例”。其
他术语的相关定义将在下文描述中给出。需要注意，本公开中提及的“第一”、“第二”
25 等概念仅用于对不同的装置、模块或单元进行区分，并非用于限定这些装置、模块或单元
所执行的功能的顺序或者相互依存关系。另外需要注意，本公开中提及的“一个”、“多
个”的修饰是示意性而非限制性的，本领域技术人员应当理解，除非在上下文另有明确指
出，否则应该理解为“一个或多个”。

本公开实施方式中的多个装置之间所交互的消息或者信息的名称仅用于说明性的目
30 的，而并不是用于对这些消息或信息的范围进行限制。

正如上文所言，相关技术中的语音合成方式通常没有考虑合成语音中的重音，导致合

成语音中没有重音，发音平淡，缺乏表现力。或者，相关技术中的语音合成方式通常随机选择输入文本中的词语进行重音添加，导致合成语音中的重音发音错误，无法得到较好的包含重音的语音合成结果。

有鉴于此，本公开提供一种语音合成方法、装置、存储介质及电子设备，以一种新的语音合成方式，使得合成语音中包括重音发音，并且使得合成语音中重音发音符合实际重音发音习惯，提高合成语音中重音发音的准确性。

图 1A 是根据本公开一示例性实施例示出的一种语音合成方法的流程图。参照图 1A，该语音合成方法包括：

步骤 101，获取标注有重音词的待合成文本。

步骤 102，将待合成文本输入语音合成模型中，以得到待合成文本对应的音频信息，该语音合成模型是通过标注有重音词的样本文本和样本文本对应的样本音频训练得到的。

通过上述方式，可以根据标注有重音词的样本文本和该样本文本对应的样本音频训练语音合成模型，训练后的语音合成模型可以根据标注有重音词的待合成文本生成包括有重音发音的音频信息。由于语音合成模型是根据标注有重音词的大量样本文本训练得到的，因此相较于相关技术中随机添加重音发音的方式，一定程度上可以保证生成的音频信息的准确性。

根据本公开的一些实施例，参照图 1B，该语音合成方法可包括采用语音合成模型用于通过如下方式对待合成文本进行处理，包括：

步骤 1021，确定待合成文本对应的音素序列；

步骤 1022，根据待合成文本中标注的重音词，确定音素级别的重音标签；

步骤 1023，根据音素序列和重音标签，生成待合成文本对应的音频信息。

通过上述方式，语音合成模型可以在将待合成文本扩展到音素级别的情况下进行语音合成处理，因此可以使得合成语音中的重音在音素级别上可控，从而进一步提升合成语音中重音发音的准确性。

为了使得本领域技术人员更加理解本公开提供的语音合成方法，下面对上述各步骤进行详细举例说明。

首先说明语音合成模型的训练过程。

根据本公开的一些实施例，可以预先获取用于训练的多个样本文本和该多个样本文本分别对应的样本音频，其中每一样本文本中标注有重音词，即每一样本文本中标注有需要重音发音的词语。

在一些实施例中，参照图 1C，样本文本中标注的重音词的确定可包括：

步骤 1031, 获取多个样本文本, 每一样本文本中包括标注有初始重音标记的重音词,

步骤 1032, 针对标注有初始重音标记的每一重音词, 若该重音词在每一个样本文本中被标注为重音词, 则对该重音词添加目标重音标记; 若该重音词在至少两个样本文本中被标注为重音词, 则在该重音词的基频大于预设基频阈值且该重音词的能量大于预设能量
5 阈值的情况下, 对该重音词添加目标重音标记,

步骤 1033, 针对每一样本文本, 将该样本中添加有目标重音标记的重音词确定为该样本文本中的重音词。

根据本公开的一些实施例, 多个样本文本可以是包括相同内容且由不同用户进行初始重音标记的样本文本, 或者可以是包括不同内容的多个文本且包括同一内容的文本由不同
10 用户进行初始重音标记, 等等, 本公开实施例对此不作限定。应当理解的是, 为了提高结果准确性, 优选的是多个样本文本为包括不同内容的多个文本且包括同一内容的文本由不同用户进行初始重音标记。

示例地, 首先可以通过自动对齐模型, 获取样本文本中每个字在样本音频中的时间边界信息, 从而获得该样本文本中每个词和每个韵律短语的时间边界信息。然后, 多个用户
15 可以根据对齐好的样本音频和样本文本, 结合听感、波形图、频谱和从样本文本中获取的语意信息, 在韵律短语级别标注重音词, 获得具有初始重音标记的多个样本文本。其中, 韵律短语是介于韵律词和语调短语之间的中等节奏组块。韵律词是一组在实际语流中联系密切的、经常联在一起发音的音节。语调短语是将几个韵律短语按照一定的句调模式连接起来, 一般对应句法上的句子。在本公开实施例中, 样本文本中的初始重音标记可以与韵
20 律短语对应, 从而得到韵律短语级别的初始重音标记, 使得重音发音更加符合常规的发音习惯。

在本公开实施例中, 或者, 在其他可能的情况下, 样本文本中的初始重音标记可以与单个字、词对应, 从而得到词语级别的重音或者单字级别的重音, 等等, 在具体实施时, 可以根据需求进行选择。

25 在获得具有初始重音标记的多个样本文本之后, 可以整合多个样本文本中的初始重音标记。具体地, 针对标注有初始重音标记的每一重音词, 若该重音词在每一个样本文本中被标注为重音词, 则说明该重音标注结果较为准确, 因此可以将对该重音词添加目标重音标记。若该重音词在至少两个样本文本中被标注为重音词, 则说明在其他样本文本中存在未将该重音词标注为重音的情况, 从而说明该重音标注结果可能存在一定偏差。在此种情
30 况下, 为了提高结果准确性, 可以进一步进行判断。比如, 考虑到音频中的重音发音的基频比非重音发音的基频更高, 且音频中的重音发音的能量比非重音发音的能量更高, 因此

可以在该重音词的基频大于预设基频阈值且该重音词的能量大于预设能量阈值的情况下，对该重音词添加目标重音标记。其中，预设基频阈值和预设能量阈值可以根据实际情况设定，本公开实施例对此不作限定。

应当理解的是，在其他可能的情况下，若重音词未包括在其他所有样本文本中，则说明该重音词只在一个样本文本中被标注为重音，因此该重音词为重音的可能性较低，从而可以不对该重音词添加目标重音标记。

通过上述方式，可以对标注有初始重音标记的样本文本进行重音标记筛选，即得到添加有目标重音标记的样本文本，从而针对每一样本文本，可以将添加有目标重音标记的重音词确定为该样本文本中的重音词，使得样本文本中的重音标记信息更加准确。

在得到标注有重音词的样本文本之后，可以根据标注有重音词的该多个样本文本以及该多个样本文本分别对应的样本音频训练语音合成模型。

在本公开实施例中，参照图 1D，语音合成模型的训练过程可包括：

步骤 1041，对样本文本对应的音素序列进行向量化，以得到样本音素向量，

步骤 1042，根据样本文本中标注的重音词，确定样本文本对应的样本重音标签，并对该样本重音标签进行向量化，以得到音素级别的样本重音标签向量，

步骤 1043，根据样本音素向量和样本重音标签向量，确定目标样本音素向量，并根据该目标样本音素向量确定样本梅尔频谱，

步骤 1044，根据该样本梅尔频谱与样本音频对应的实际梅尔频谱计算损失函数，并通过该损失函数调整语音合成模型的参数。

应当理解的是，音素是根据语音的自然属性划分出来的最小语音单位，分为元音与辅音两大类。对于中文来说，音素包括声母（声母是使用在韵母前面的辅音，跟韵母一齐构成的一个完整的音节）和韵母（即元音）。对于英文来说，音素包括元音和辅音。在本公开实施例中，语音合成模型的训练阶段首先对样本文本对应的音素序列进行向量化，以得到样本音素向量，可以在后续过程中合成带有音素级别重音的语音，使得合成语音中的重音在音素级别上可控，从而进一步提升合成语音中重音发音的准确性。其中，对样本文本对应的音素序列进行向量化，以得到样本音素向量的过程与相关技术中的向量转换方式类似，这里不再赘述。

示例地，根据样本文本中标注的重音词，确定样本文本对应的样本重音标签可以根据样本文本中标注的重音词，生成通过 0 和 1 表示的重音序列。其中，0 表示未标注重音，1 表示标注有重音。然后，可以将该样本重音标签进行向量化，以得到样本重音标签向量。在具体应用时，可以先确定样本文本对应的音素序列，然后根据样本文本中标注的重音词，

在该样本文本对应的音素序列中进行重音标注，从而得到该样本文本对应的音素级别的样本重音标签，进而对该样本重音标签进行向量化，以得到音素级别的样本重音标签向量。其中，对样本重音标签进行向量化，以得到音素级别的样本重音标签向量的方式与相关技术中的向量转换方式类似，这里也不再赘述。

5 在得到样本音素向量和样本重音标签向量之后，可以根据样本音素向量和样本重音标签向量，确定目标样本音素向量，从而根据目标样本音素向量确定样本梅尔频谱。其中，考虑到样本音素向量和样本重音标签向量表征的是相互独立的两个信息，因此可以通过拼接样本音素向量和样本重音标签向量的方式得到目标样本音素向量，而不是通过将样本音素向量和样本重音标签向量相加的方式得到目标样本音素向量，从而避免破坏样本音素向
10 量和样本重音标签向量之间的内容独立性，保证语音合成模型输出结果的准确性。

 在一些实施例，根据目标样本音素向量确定样本梅尔频谱可以是：将目标样本音素向量输入编码器，然后将编码器输出的向量输入译码器，以得到样本梅尔频谱；其中，编码器用于确定输入向量对应的音素序列中每一音素的发音信息，译码器用于根据输入向量对应的每一音素的发音信息进行转换处理，以得到每一音素对应的梅尔频谱。或者，还可
15 以通过自动对齐模型确定编码器输出的向量对应的帧级别的向量，然后将该帧级别的向量输入译码器，以得到样本梅尔频谱，其中，自动对齐模型用于将目标样本音素向量对应的样本文本中音素级别的发音信息与目标样本音素向量对应的样本音频中每一音素的帧时间一一对应，以提高模型训练效果，进而提高模型合成语音中重音发音的准确性。

 示例地，语音合成模型可以为端到端的语音合成 Tacotron 模型，相应地，编码器可
20 以是 Tacotron 模型中的编码器，译码器可以是 Tacotron 模型中的译码器。例如，语音合成模型如图 2 所示，在语音合成模型的训练阶段，向量化音素序列（比如样本音素向量）和向量化重音标签（比如样本重音标签向量）拼接得到目标样本音素向量之后，可以将该目标样本音素向量输入编码器（Encoder），以得到该目标样本音素向量对应的音素序列中每一音素的发音信息。比如对于目标样本音素向量对应的音素序列中包括音素“jin”，需
25 要知道该音素的发音同“今”。然后，可以通过自动对齐模型实现音素级别和帧级别的对齐，得到编码器输出的向量对应的帧级别的目标样本向量。接着，可以将目标样本向量输入译码器（Decoder），以使译码器根据目标样本向量对应的音素序列中每一音素的发音信息进行转换处理，从而得到每一音素对应的样本梅尔频谱（Mel 谱）。

 在另一种可能的方式中，参照图 3，也可以先将样本音素向量输入编码器，然后将编
30 码器输出的向量与样本重音标签向量进行拼接，以得到目标样本音素向量，从而根据该目标样本音素向量确定样本梅尔频谱。在实际应用时，可以根据需求选择将拼接处理设置在

编码器之前或设置在编码器之后，本公开实施例对此不作限定。

在得到样本梅尔频谱之后，可以根据样本梅尔频谱与样本音频对应的实际梅尔频谱计算损失函数，并通过该损失函数调整语音合成模型的参数。比如，可以根据样本梅尔频谱与实际梅尔频谱计算 MSE 损失函数，然后通过该 MSE 损失函数调整语音合成模型的参数。

5 或者，还可以通过 Adam 优化器进行模型优化，从而保证训练后语音合成模型的输出结果的准确性。

通过上述方式训练得到语音合成模型之后，则可以通过该语音合成模型对标注有重音词的待合成文本进行语音合成。也即是说，针对标注有重音词的待合成文本，该语音合成模型可以输出与该待合成文本中对应的音频信息，且该音频信息中带有与该待合成文本中
10 标注的重音词所对应的重音发音，从而可以解决相关技术中合成语音无重音，减少重音发音错误的问题，提高合成语音中重音发音的准确性。

示例地，用户可以根据通常的重音发音习惯标注待合成文本中的重音词，比如待合成文本为“今天天气真好啊”，根据通常的重音发音习惯，可以将该待合成文本中的“真好”标注为重音词。然后用户可以将标注有重音词的待合成文本输入电子设备进行语音合成。
15 相应地，电子设备可以响应于用户输入待合成文本的操作，获取标注有重音词的该待合成文本进行语音合成。其中，本公开实施例对于待合成文本的具体内容和内容长度不作限定，比如待合成文本可以是单句，或者也可以是多个句子，等等。

在获取到标注有重音词的待合成文本之后，电子设备可以将该待合成文本输入预训练的语音合成模型。示例地，该语音合成模型可以先确定待合成文本对应的音素序列，从而
20 在后续过程中可以在音素级别上合成带重音的语音，使得合成语音中的重音在音素级别上可控，进一步提升合成语音中重音发音的准确性。

在确定待合成文本对应的音素序列的同时或之后，还可以根据待合成文本中标注的重音词，确定音素级别的重音标签。示例地，该重音标签可以是表示为 0 和 1 的序列，其中 0 表示待合成文本中对应音素未标注重音，1 表示待合成文本中对应音素标注有重音。在
25 具体应用时，可以先确定待合成文本对应的音素序列，然后根据待合成文本中标注有的重音词，在该音素序列中进行重音标注，从而得到音素级别的重音标签。

在得到待合成文本对应的音素序列和重音标签之后，可以根据该音素序列和该重音标签，生成待合成文本对应的音频信息。示例地，语音合成模型可以将待合成文本对应的音素序列进行向量化，以得到音素向量，并将重音标签进行向量化，以得到重音标签向量，
30 然后根据该音素向量和该重音标签向量，确定目标音素向量，并根据目标音素向量确定梅尔频谱，最后将梅尔频谱输入声码器，以得到待合成文本对应的音频信息。

应当理解的是，将待合成文本对应的音素序列进行向量化，以得到音素向量的过程、以及将待合成文本对应的重音标签进行向量化，以得到重音标签向量的过程与相关技术中的向量转换方式类似，这里不再赘述。

示例地，考虑到音素向量和重音标签向量表征的是相互独立的两个信息，因此可以通过拼接音素向量和重音标签向量的方式得到目标音素向量，而不是通过将音素向量和重音标签向量相加的方式得到目标音素向量，从而避免破坏音素向量和重音标签向量之间的内容独立性，保证后续语音合成结果的准确性。

在得到目标音素向量之后，可以根据该目标音素向量确定梅尔频谱（Mel 谱）。示例地，可以将目标音素向量输入编码器，并将编码器输出的向量输入译码器，以得到对应的梅尔频谱，其中，编码器用于确定输入向量对应的音素序列中每一音素的发音信息，译码器用于根据输入向量对应的每一音素的发音信息进行转换处理，以得到每一音素对应的梅尔频谱。

例如，参照图 2 所示，本公开实施例中的语音合成模型可以包括编码器（Encoder）和译码器（Decoder）。相应地，在拼接得到目标音素向量后，可以将该目标音素向量输入编码器，以得到该目标音素向量对应的音素序列中每一音素的发音信息，比如对于音素“jin”，需要知道该音素的发音同“今”。然后将发音信息输入译码器中，以通过译码器根据目标音素向量对应的音素序列中每一音素的发音信息进行转换处理，得到每一音素对应的梅尔频谱。

或者，在其他可能的方式中，可以将音素向量输入编码器，并根据编码器输出的向量和重音标签向量，确定目标音素向量。相应地，可以将目标音素向量输入译码器，以得到对应的梅尔频谱。例如，参照图 3，先将音素向量输入编码器，然后将编码器输出的向量与重音标签向量拼接，以得到目标音素向量，从而根据目标音素向量确定梅尔频谱。

在确定梅尔频谱之后，可以将该梅尔频谱输入声码器，以得到待合成文本对应的音频信息。应当理解的是，本公开实施例对于声码器的类型不作限定，也即是说将梅尔频谱输入任意声码器即可得到带有重音的音频信息，并且该音频信息中的重音与待合成文本中标注的重音词对应，从而解决相关技术中合成语音没有重音或者由于随机指定重音而导致重音发音错误的问题，提高合成语音中的重音发音准确性。

根据本公开的实施例，本公开还提供一种语音合成装置，该语音合成装置可以通过软件、硬件或者两者结合的方式成为电子设备的部分或全部。参照图 4，该语音合成装置 400 包括：

获取模块 401，用于获取标注有重音词的待合成文本；

合成模块 402，用于将所述待合成文本输入语音合成模型中，以得到所述待合成文本对应的音频信息，所述语音合成模型是通过标注有重音词的样本文本和所述样本文本对应的样本音频训练得到的，所述语音合成模型用于通过如下模块对所述待合成文本进行处理：

第一确定子模块 4021，用于确定所述待合成文本对应的音素序列；

- 5 第二确定子模块 4022，用于根据所述待合成文本中标注的所述重音词，确定音素级别的重音标签；

生成子模块 4023，用于根据所述音素序列和所述重音标签，生成所述待合成文本对应的音频信息。

在一些实施例中，所述生成子模块 4023 用于：

- 10 将所述待合成文本对应的所述音素序列进行向量化，以得到音素向量，并将所述重音标签进行向量化，以得到重音标签向量；

根据所述音素向量和所述重音标签向量，确定目标音素向量；

根据所述目标音素向量确定梅尔频谱；

将所述梅尔频谱输入声码器，以得到所述待合成文本对应的音频信息。

- 15 在一些实施例中，所述生成子模块 4023 用于：

将所述目标音素向量输入编码器，并将所述编码器输出的向量输入译码器，以得到对应的梅尔频谱，其中，所述编码器用于确定输入向量对应的音素序列中每一音素的发音信息，所述译码器用于根据输入向量对应的每一音素的发音信息进行转换处理，以得到所述每一音素对应的梅尔频谱。

- 20 在一些实施例中，所述生成子模块 4023 用于：

将所述音素向量输入编码器，并根据所述编码器输出的向量和所述重音标签向量，确定所述目标音素向量；

将所述目标音素向量输入译码器，以得到所述梅尔频谱；

- 25 其中，所述编码器用于确定输入向量对应的音素序列中每一音素的发音信息，所述译码器用于根据输入向量对应的每一音素的发音信息进行转换处理，以得到所述每一音素对应的梅尔频谱。

在一些实施例中，所述装置 400 还可包括重音词确定模块 403，该重音词确定模块 403 可包括以下模块：

- 30 样本获取模块 4031，用于获取多个样本文本，每一所述样本文本中包括标注有初始重音标记的重音词；

添加模块 4032，用于针对标注有所述初始重音标记的每一所述重音词，若该重音词

包括在每一所述样本文本中，则将对该重音词添加目标重音标记；若该重音词包括在至少两个所述样本文本中，则在该重音词的基频大于预设基频阈值且该重音词的能量大于预设能量阈值的情况下，对该重音词添加目标重音标记；

5 标注模块 4033，用于针对每一所述样本文本，将该样本文本中添加有所述目标重音标记的重音词确定为该样本文本中的重音词。

在一些实施例中，所述装置 400 还可包括语音合成模型确定模块 404，语音合成模型确定模块 404 包括以下模块：

第一训练模块 4041，用于对所述样本文本对应的音素序列进行向量化，以得到样本音素向量；

10 第二训练模块 4042，用于根据所述样本文本中标注的所述重音词，确定所述样本文本对应的样本重音标签，并对所述样本重音标签进行向量化，以得到样本重音标签向量；

第三训练模块 4043，用于根据所述样本音素向量与所述样本重音标签向量，确定目标样本音素向量，并根据所述目标样本音素向量确定样本梅尔频谱；

15 第四训练模块 4044，用于根据所述样本梅尔频谱与所述样本音频对应的实际梅尔频谱计算损失函数，并通过所述损失函数调整所述语音合成模型的参数。

关于上述实施例中的装置，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。应注意，上述各个模块的划分并非限制具体的实现方式，上述各个模块例如可以以软件、硬件或者软硬件结合的方式来实现。在实际实现时，上述各个模块可被实现为独立的物理实体，或者也可由单个实体（例如，
20 处理器（CPU 或 DSP 等）、集成电路等）来实现。需要注意的是，尽管图 4 中将各个模块示为分立的模块，但是这些模块中的一个或多个也可以合并为一个模块，或者拆分为多个模块。此外，上述重音词确定模块和语音合成模型确定模块在附图中用虚线示出指示这些模块不必被包含在语音合成装置中，其可以在语音合成装置之外实现或者由语音合成装置之外的其它设备实现并且将结果告知语音合成装置。或者，上述重音词确定模块和语音合成模型确定模块在附图中用虚线示出指示这些模块可以并不实际存在，而它们所实现的操作/功能可由语音合成装置本身来实现。
25

根据本公开的一些实施例，本公开还提供一种计算机可读介质，其上存储有计算机程序，该程序被处理装置执行时实现上述任一语音合成方法的步骤。

根据本公开的一些实施例，本公开还提供一种电子设备，包括：

30 存储装置，其上存储有计算机程序；

处理装置，用于执行所述存储装置中的所述计算机程序，以实现上述任一语音合成方

根据本公开的一些实施例，本公开还提供一种计算机程序产品，包括指令，所述指令在由计算机执行时使得计算机实现上述任一语音合成方法的步骤。

下面参考图 5，其示出了适于用来实现本公开实施例的电子设备 500 的结构示意图。

5 本公开实施例中的终端设备可以包括但不限于诸如移动电话、笔记本电脑、数字广播接收器、PDA（个人数字助理）、PAD（平板电脑）、PMP（便携式多媒体播放器）、车载终端（例如车载导航终端）等等的移动终端以及诸如数字 TV、台式计算机等等的固定终端。图 5 示出的电子设备仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

如图 5 所示，电子设备 500 可以包括处理装置（例如中央处理器、图形处理器等）501，
10 其可以根据存储在只读存储器（ROM）502 中的程序或者从存储装置 508 加载到随机访问存储器（RAM）503 中的程序而执行各种适当的动作和处理。在 RAM 503 中，还存储有电子设备 500 操作所需的各种程序和数据。处理装置 501、ROM 502 以及 RAM 503 通过总线 504 彼此相连。输入/输出（I/O）接口 505 也连接至总线 504。

通常，以下装置可以连接至 I/O 接口 505：包括例如触摸屏、触摸板、键盘、鼠标、
15 摄像头、麦克风、加速度计、陀螺仪等的输入装置 506；包括例如液晶显示器（LCD）、扬声器、振动器等的输出装置 507；包括例如磁带、硬盘等的存储装置 508；以及通信装置 509。通信装置 509 可以允许电子设备 500 与其他设备进行无线或有线通信以交换数据。虽然图 5 示出了具有各种装置的电子设备 500，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。

20 特别地，根据本公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在非暂态计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信装置 509 从网络上被下载和安装，或者从存储装置 508 被安装，或者从 ROM 502 被安装。在该计算机程序被处理装置 501 执行时，执行本
25 公开实施例的方法中限定的上述功能。

需要说明的是，本公开上述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、
30 便携式计算机磁盘、硬盘、随机访问存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑磁盘只读存储器（CD-ROM）、光存储器件、

磁存储器件、或者上述的任意合适的组合。在本公开中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中，计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：电线、光缆、RF（射频）等等，或者上述的任意合适的组合。

在一些实施方式中，可以利用诸如 HTTP（HyperText Transfer Protocol，超文本传输协议）之类的任何当前已知或未来研发的网络协议进行通信，并且可以与任意形式或介质的数字数据通信（例如，通信网络）互连。通信网络的示例包括局域网（“LAN”），广域网（“WAN”），网际网（例如，互联网）以及端对端网络（例如，ad hoc 端对端网络），以及任何当前已知或未来研发的网络。

上述计算机可读介质可以是上述电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。

上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该电子设备执行时，使得该电子设备：获取标注有重音词的待合成文本；将所述待合成文本输入语音合成模型中，以得到所述待合成文本对应的音频信息，所述语音合成模型是通过标注有重音词的样本文本和所述样本文本对应的样本音频训练得到的，所述语音合成模型用于通过如下方式对所述待合成文本进行处理：确定所述待合成文本对应的音素序列；根据所述待合成文本中标注的所述重音词，确定音素级别的重音标签；根据所述音素序列和所述重音标签，生成所述待合成文本对应的音频信息。

可以以一种或多种程序设计语言或其组合来编写用于执行本公开的操作的计算机程序代码，上述程序设计语言包括但不限于面向对象的程序设计语言——诸如 Java、Smalltalk、C++，还包括常规的过程式程序设计语言——诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网（LAN）或广域网（WAN）——连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。

附图中的流程图和框图，图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

描述于本公开实施例中所涉及到的模块可以通过软件的方式实现，也可以通过硬件的方式来实现。其中，模块的名称在某种情况下并不构成对该模块本身的限定。

本文中以上描述的功能可以至少部分地由一个或多个硬件逻辑部件来执行。例如，非限制性地，可以使用的示范类型的硬件逻辑部件包括：现场可编程门阵列（FPGA）、专用集成电路（ASIC）、专用标准产品（ASSP）、片上系统（SOC）、复杂可编程逻辑设备（CPLD）等等。

在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦除可编程只读存储器（EPROM 或快闪存储器）、光纤、便捷式紧凑盘只读存储器（CD-ROM）、光学储存设备、磁储存设备、或上述内容的任何合适组合。

根据本公开的一个或多个实施例，示例 1 提供了一种语音合成方法，所述包括：

获取标注有重音词的待合成文本；

将所述待合成文本输入语音合成模型中，以得到所述待合成文本对应的音频信息，所述语音合成模型是通过标注有重音词的样本文本和所述样本文本对应的样本音频训练得到的，所述语音合成模型用于通过如下方式对所述待合成文本进行处理：

确定所述待合成文本对应的音素序列；

根据所述待合成文本中标注的所述重音词，确定音素级别的重音标签；

根据所述音素序列和所述重音标签，生成所述待合成文本对应的音频信息。

根据本公开的一个或多个实施例，示例性实施例 2 提供了示例性实施例 1 的方法，所述根据所述音素序列和所述重音标签，生成所述待合成文本对应的音频信息，包括：

将所述待合成文本对应的所述音素序列进行向量化，以得到音素向量，并将所述重音标签进行向量化，以得到重音标签向量；

5 根据所述音素向量和所述重音标签向量，确定目标音素向量；

根据所述目标音素向量确定梅尔频谱；

将所述梅尔频谱输入声码器，以得到所述待合成文本对应的音频信息。

根据本公开的一个或多个实施例，示例性实施例 3 提供了示例性实施例 2 的方法，所述根据所述目标音素向量，确定梅尔频谱，包括：

10 将所述目标音素向量输入编码器，并将所述编码器输出的向量输入译码器，以得到对应的梅尔频谱，其中，所述编码器用于确定输入向量对应的音素序列中每一音素的发音信息，所述译码器用于根据输入向量对应的每一音素的发音信息进行转换处理，以得到所述每一音素对应的梅尔频谱。

根据本公开的一个或多个实施例，示例性实施例 4 提供了示例性实施例 2 的方法，所述根据所述音素向量和所述重音标签向量，确定目标音素向量，包括：

15 将所述音素向量输入编码器，并根据所述编码器输出的向量和所述重音标签向量，确定所述目标音素向量；

所述根据所述目标音素向量确定梅尔频谱，包括：

将所述目标音素向量输入译码器，以得到所述梅尔频谱；

20 其中，所述编码器用于确定输入向量对应的音素序列中每一音素的发音信息，所述译码器用于根据输入向量对应的每一音素的发音信息进行转换处理，以得到所述每一音素对应的梅尔频谱。

根据本公开的一个或多个实施例，示例性实施例 5 提供了示例性实施例 1-4 任一项的方法，所述样本文本中标注的所述重音词是通过如下方式确定的：

25 获取多个样本文本，每一所述样本文本中包括标注有初始重音标记的重音词；

针对标注有所述初始重音标记的每一所述重音词，若该重音词在每一个所述样本文本中被标注为重音词，则对该重音词添加目标重音标记；若该重音词在至少两个所述样本文本中被标注为重音词，则在该重音词的基频大于预设基频阈值且该重音词的能量大于预设能量阈值的情况下，对该重音词添加目标重音标记；

30 针对每一所述样本文本，将该样本文本中添加有所述目标重音标记的重音词确定为该样本文本中的重音词。

根据本公开的一个或多个实施例，示例性实施例 6 提供了示例性实施例 5 的方法，所述语音合成模型是通过如下方式训练得到的：

对所述样本文本对应的音素序列进行向量化，以得到样本音素向量；

根据所述样本文本中标注的所述重音词，确定所述样本文本对应的样本重音标签，并对所述样本重音标签进行向量化，以得到音素级别的样本重音标签向量；

根据所述样本音素向量与所述样本重音标签向量，以得到目标音素向量，并根据所述目标音素向量确定样本梅尔频谱；

根据所述样本梅尔频谱与所述样本音频对应的实际梅尔频谱计算损失函数，并通过所述损失函数调整所述语音合成模型的参数。

10 根据本公开的一个或多个实施例，示例性实施例 7 提供了一种语音合成装置，所述装置包括：

获取模块，用于获取标注有重音词的待合成文本；

合成模块，用于将所述待合成文本输入语音合成模型中，以得到所述待合成文本对应的音频信息，所述语音合成模型是通过标注有重音词的样本文本和所述样本文本对应的样本音频训练得到的，所述语音合成模型用于通过如下模块对所述待合成文本进行处理：

第一确定子模块，用于确定所述待合成文本对应的音素序列；

第二确定子模块，用于根据所述待合成文本中标注的所述重音词，确定音素级别的重音标签；

生成子模块，用于根据所述音素序列和所述重音标签，生成所述待合成文本对应的音频信息。

20 根据本公开的一个或多个实施例，示例性实施例 8 提供了示例性实施例 7 的装置，所述生成子模块用于：

将所述待合成文本对应的所述音素序列进行向量化，以得到音素向量，并将所述重音标签进行向量化，以得到重音标签向量；

25 拼接所述音素向量和所述重音标签向量，以得到目标音素向量；

根据所述目标音素向量确定梅尔频谱；

将所述梅尔频谱输入声码器，以得到所述待合成文本对应的音频信息。

根据本公开的一个或多个实施例，示例性实施例 9 提供了示例性实施例 8 的装置，所述生成子模块用于：

30 将所述目标音素向量输入编码器，并将所述编码器输出的向量输入译码器，以得到对应的梅尔频谱，其中，所述编码器用于确定输入向量对应的音素序列中每一音素的发音信

息, 所述译码器用于根据输入向量对应的每一音素的发音信息进行转换处理, 以得到所述每一音素对应的梅尔频谱。

根据本公开的一个或多个实施例, 示例性实施例 10 提供了示例性实施例 8 的装置, 所述生成子模块用于:

- 5 将所述音素向量输入编码器, 并根据所述编码器输出的向量和所述重音标签向量, 确定所述目标音素向量;

 将所述目标音素向量输入译码器, 以得到所述梅尔频谱;

- 其中, 所述编码器用于确定输入向量对应的音素序列中每一音素的发音信息, 所述译码器用于根据输入向量对应的每一音素的发音信息进行转换处理, 以得到所述每一音素对
10 应的梅尔频谱。

根据本公开的一个或多个实施例, 示例性实施例 11 提供了示例性实施例 7 至 10 任一
项的装置, 还包括用于确定所述样本文本中标注的所述重音词的以下模块:

 样本获取模块, 用于获取多个样本文本, 每一所述样本文本中包括标注有初始重音标
记的重音词;

- 15 添加模块, 用于针对标注有所述初始重音标记的每一所述重音词, 若该重音词包括在
每一所述样本文本中, 则将该重音词添加目标重音标记; 若该重音词包括在至少两个所
述样本文本中, 则在该重音词的基频大于预设基频阈值且该重音词的能量大于预设能量阈
值的情况下, 对该重音词添加目标重音标记;

- 标注模块, 用于针对每一所述样本文本, 将该样本文本中添加有所述目标重音标记的
20 重音词确定为该样本文本中的重音词。

根据本公开的一个或多个实施例, 示例性实施例 12 提供了示例性实施例 11 的装置,
还包括用于训练所述语音合成模型的以下模块:

 第一训练模块, 用于对所述样本文本对应的音素序列进行向量化, 以得到样本音素向
量;

- 25 第二训练模块, 用于根据所述样本文本中标注的所述重音词, 确定所述样本文本对应
的样本重音标签, 并对所述样本重音标签进行向量化, 以得到样本重音标签向量;

 第三训练模块, 用于拼接所述样本音素向量与所述样本重音标签向量, 以得到目标样
本音素向量, 并根据所述目标样本音素向量确定样本梅尔频谱;

- 第四训练模块, 用于根据所述样本梅尔频谱与所述样本音频对应的实际梅尔频谱计算
30 损失函数, 并通过所述损失函数调整所述语音合成模型的参数。

根据本公开的一个或多个实施例, 示例性实施例 13 提供了一种计算机可读介质, 其

上存储有计算机程序，该程序被处理装置执行时实现示例性实施例 1 至 6 中任一语音合成方法的步骤。

根据本公开的一个或多个实施例，示例性实施例 14 提供了一种电子设备，包括：

存储装置，其上存储有计算机程序；

- 5 处理装置，用于执行所述存储装置中的所述计算机程序，以实现示例性实施例 1 至 6 中任一语音合成方法的步骤。

以上描述仅为本公开的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本公开中所涉及的公开范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述公开构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开中公开的（但不限于）具有类似功能的技术特征进行互相替换而形成的技术方案。

10

此外，虽然采用特定次序描绘了各操作，但是这不应理解为要求这些操作以所示出的特定次序或以顺序次序执行来执行。在一定环境下，多任务和并行处理可能是有利的。同样地，虽然在上面论述中包含了若干具体实现细节，但是这些不应被解释为对本公开的范围的限制。在单独的实施例的上下文中描述的某些特征还可以组合地实现在单个实施例中。相反地，在单个实施例的上下文中描述的各种特征也可以单独地或以任何合适的子组合的方式实现在多个实施例中。

15

尽管已经采用特定于结构特征和/或方法逻辑动作的语言描述了本主题，但是应当理解所附权利要求书中所限定的主题未必局限于上面描述的特定特征或动作。相反，上面所描述的特定特征和动作仅仅是实现权利要求书的示例形式。关于上述实施例中的装置，其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

20

权 利 要 求

1、一种语音合成方法，所述方法包括：

获取标注有重音词的待合成文本；

将所述待合成文本输入语音合成模型中，以得到所述待合成文本对应的音频信息，所述语音合成模型是通过标注有重音词的样本文本和所述样本文本对应的样本音频训练得到的。

2、根据权利要求 1 所述的方法，其中，所述语音合成模型用于通过如下方式对所述待合成文本进行处理：

确定所述待合成文本对应的音素序列；

根据所述待合成文本中标注的所述重音词，确定音素级别的重音标签；

根据所述音素序列和所述重音标签，生成所述待合成文本对应的音频信息。

3、根据权利要求 2 所述的方法，其中，所述根据所述音素序列和所述重音标签，生成所述待合成文本对应的音频信息，包括：

将所述待合成文本对应的所述音素序列进行向量化，以得到音素向量，并将所述重音标签进行向量化，以得到重音标签向量；

根据所述音素向量和所述重音标签向量，确定目标音素向量；

根据所述目标音素向量确定梅尔频谱；

将所述梅尔频谱输入声码器，以得到所述待合成文本对应的音频信息。

4、根据权利要求 3 所述的方法，其中，所述根据所述目标音素向量确定梅尔频谱，包括：

将所述目标音素向量输入编码器，并将所述编码器输出的向量输入译码器，以得到对应的梅尔频谱，其中，所述编码器用于确定输入向量对应的音素序列中每一音素的发音信息，所述译码器用于根据输入向量对应的每一音素的发音信息进行转换处理，以得到所述每一音素对应的梅尔频谱。

5、根据权利要求 4 所述的方法，其中，通过拼接音素向量和重音标签向量得到目标音素向量。

6、根据权利要求 3 所述的方法，其中，所述根据所述音素向量和所述重音标签向量，确定目标音素向量，包括：

将所述音素向量输入编码器，并根据所述编码器输出的向量和所述重音标签向量，确定所述目标音素向量；

所述根据所述目标音素向量确定梅尔频谱，包括：

将所述目标音素向量输入译码器，以得到所述梅尔频谱；

其中，所述编码器用于确定输入向量对应的音素序列中每一音素的发音信息，所述译码器用于根据输入向量对应的每一音素的发音信息进行转换处理，以得到所述每一音素对应的梅尔频谱。

7、根据权利要求 6 所述的方法，其中，通过拼接所述编码器输出的向量和所述重音标签向量得到所述目标音素向量。

8、根据权利要求 1-7 任一项所述的方法，其中，所述样本文本中标注的所述重音词是通过如下方式确定的：

获取多个样本文本，每一所述样本文本中包括标注有初始重音标记的重音词；

针对标注有所述初始重音标记的每一所述重音词，若该重音词在每一个所述样本文本中被标注为重音词，则对该重音词添加目标重音标记；若该重音词在至少两个所述样本文本中被标注为重音词，则在该重音词的基频大于预设基频阈值且该重音词的能量大于预设能量阈值的情况下，对该重音词添加目标重音标记；

针对每一所述样本文本，将该样本文本中添加有所述目标重音标记的重音词确定为该样本文本中的重音词。

9、根据权利要求 8 所述的方法，其中，多个样本文本为包括不同内容的多个文本且包括同一内容的文本由不同用户进行初始重音标记。

10、根据权利要求 8 所述的方法，其中，样本文本中的初始重音标记与韵律短语对应。

11、根据权利要求 1-10 中任一项所述的方法，其中，所述语音合成模型是通过如下方式训练得到的：

对所述样本文本对应的音素序列进行向量化，以得到样本音素向量；

根据所述样本文本中标注的所述重音词，确定所述样本文本对应的样本重音标签，并对所述样本重音标签进行向量化，以得到音素级别的样本重音标签向量；

根据所述样本音素向量和所述样本重音标签向量，确定目标样本音素向量，并根据所述目标样本音素向量确定样本梅尔频谱；

根据所述样本梅尔频谱与所述样本音频对应的实际梅尔频谱计算损失函数，并通过所述损失函数调整所述语音合成模型的参数。

12、一种语音合成装置，所述装置包括：

获取模块，用于获取标注有重音词的待合成文本；

合成模块，用于将所述待合成文本输入语音合成模型中，以得到所述待合成文本对应的音频信息，所述语音合成模型是通过标注有重音词的样本文本和所述样本文本对应的样本音频训练得到的。

13、根据权利要求 12 所述的语音合成装置，其中，所述语音合成模型包括：

第一确定子模块，用于确定所述待合成文本对应的音素序列；

第二确定子模块，用于根据所述待合成文本中标注的所述重音词，确定确定音素级别的重音标签；

生成子模块，用于根据所述音素序列和所述重音标签，生成所述待合成文本对应的音频信息。

14、根据权利要求 13 所述的装置，其中，所述生成子模块用于：

将所述待合成文本对应的所述音素序列进行向量化，以得到音素向量，并将所述重音标签进行向量化，以得到重音标签向量；

根据所述音素向量和所述重音标签向量，确定目标音素向量；

根据所述目标音素向量确定梅尔频谱；

将所述梅尔频谱输入声码器，以得到所述待合成文本对应的音频信息。

15、根据权利要求 14 所述的装置，其中，所述生成子模块用于：

将所述目标音素向量输入编码器，并将所述编码器输出的向量输入译码器，以得到对应的梅尔频谱，其中，所述编码器用于确定输入向量对应的音素序列中每一音素的发音信息，所述译码器用于根据输入向量对应的每一音素的发音信息进行转换处理，以得到所述

16、根据权利要求 15 所述的装置，其中，通过拼接音素向量和重音标签向量得到目标音素向量。

17、根据权利要求 14 所述的装置，其中，所述生成子模块用于：

将所述音素向量输入编码器，并根据所述编码器输出的向量和所述重音标签向量，确定所述目标音素向量；

将所述目标音素向量输入译码器，以得到所述梅尔频谱；

其中，所述编码器用于确定输入向量对应的音素序列中每一音素的发音信息，所述译码器用于根据输入向量对应的每一音素的发音信息进行转换处理，以得到所述每一音素对应的梅尔频谱。

18、根据权利要求 17 所述的装置，其中，通过拼接所述编码器输出的向量和所述重音标签向量得到所述目标音素向量。

19、根据权利要求 12-18 中任一项所述的装置，还包括重音词确定模块，所述重音词确定模块包括：

样本获取模块，用于获取多个样本文本，每一所述样本文本中包括标注有初始重音标记的重音词；

添加模块，用于针对标注有所述初始重音标记的每一所述重音词，若该重音词包括在每一所述样本文本中，则将该重音词添加目标重音标记；若该重音词包括在至少两个所述样本文本中，则在该重音词的基频大于预设基频阈值且该重音词的能量大于预设能量阈值的情况下，对该重音词添加目标重音标记；

标注模块，用于针对每一所述样本文本，将该样本文本中添加有所述目标重音标记的重音词确定为该样本文本中的重音词。

20、根据权利要求 19 所述的装置，其中，多个样本文本为包括不同内容的多个文本且包括同一内容的文本由不同用户进行初始重音标记。

21、根据权利要求 19 所述的装置，其中，样本文本中的初始重音标记与韵律短语对应。

22、根据权利要求 12-21 中任一项所述的装置，还包括与语音合成模型训练模块，所述语音合成模型训练模块包括：

第一训练模块，用于对所述样本文本对应的音素序列进行向量化，以得到样本音素向量；

第二训练模块，用于根据所述样本文本中标注的所述重音词，确定所述样本文本对应的样本重音标签，并对所述样本重音标签进行向量化，以得到样本重音标签向量；

第三训练模块，用于拼接所述样本音素向量与所述样本重音标签向量，以得到目标样本音素向量，并根据所述目标样本音素向量确定样本梅尔频谱；

第四训练模块，用于根据所述样本梅尔频谱与所述样本音频对应的实际梅尔频谱计算损失函数，并通过所述损失函数调整所述语音合成模型的参数。

23、一种计算机可读介质，其上存储有计算机程序，其特征在于，该程序被处理装置执行时实现根据权利要求 1-11 中任一项所述方法的步骤。

24、一种电子设备，其特征在于，包括：

存储装置，其上存储有计算机程序；

处理装置，用于执行所述存储装置中的所述计算机程序，以实现根据权利要求 1-11 中任一项所述方法的步骤。

25、一种计算机程序产品，包括指令，所述指令在由计算机执行时使得计算机实现根据权利要求 1-11 中任一项所述方法的步骤。

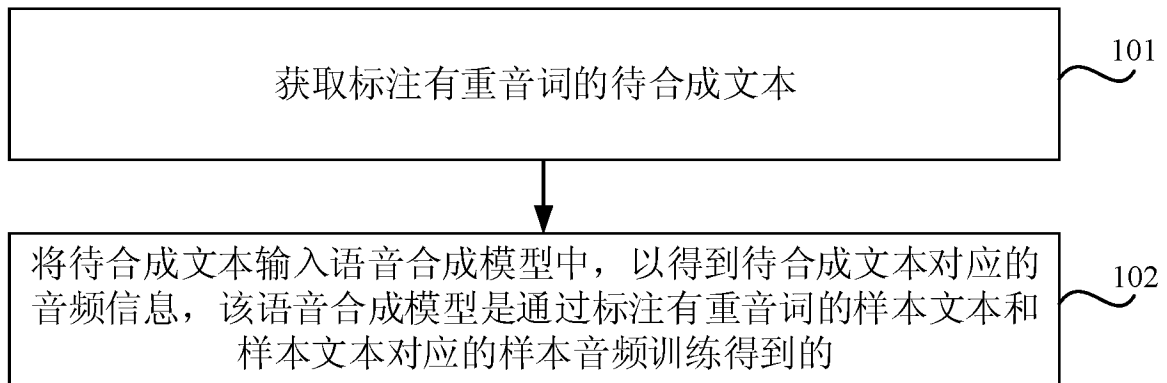


图 1A

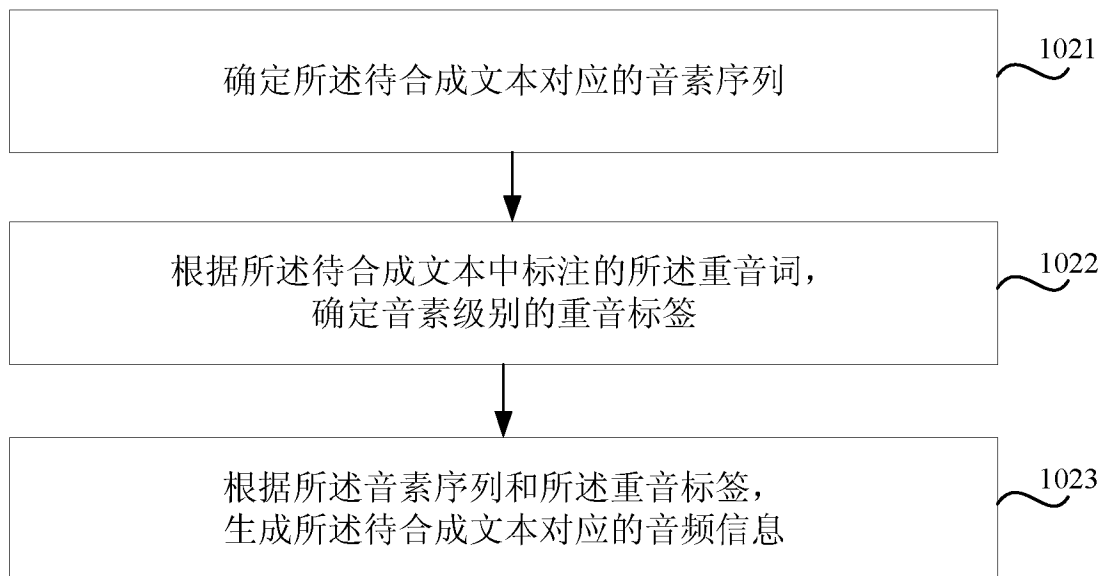


图 1B

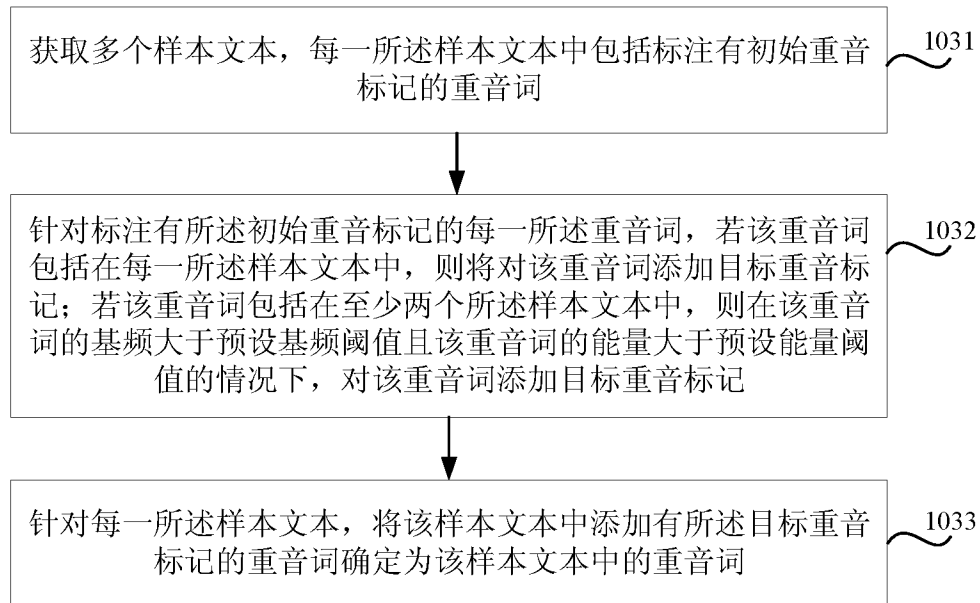


图 1C

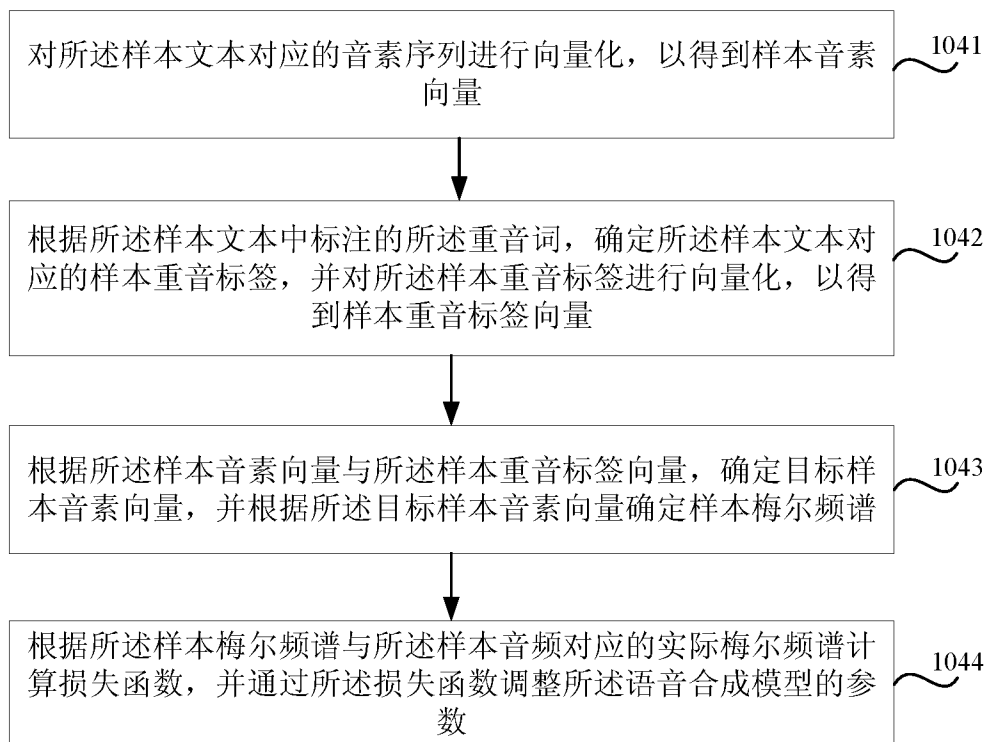


图 1D

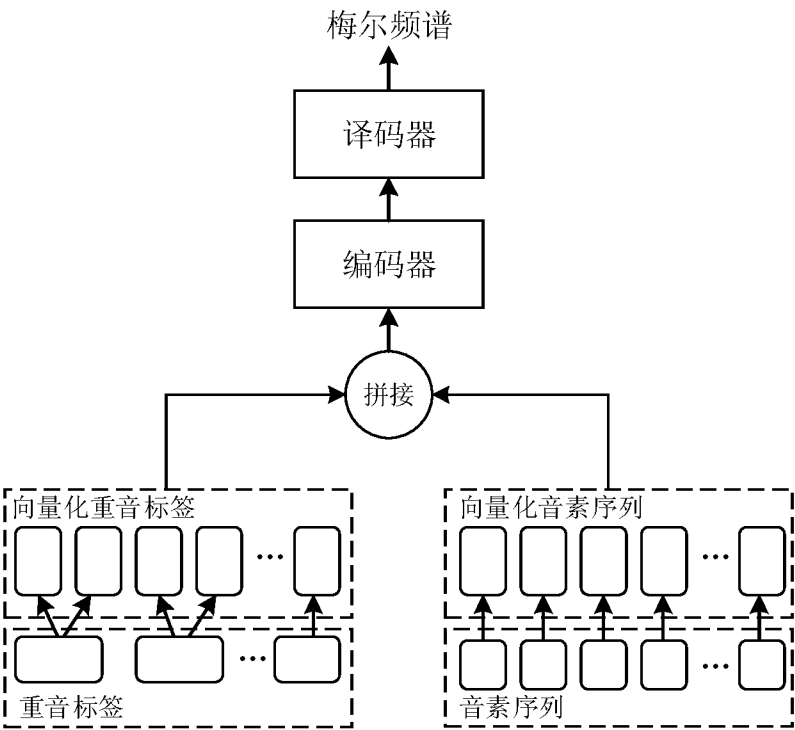


图 2

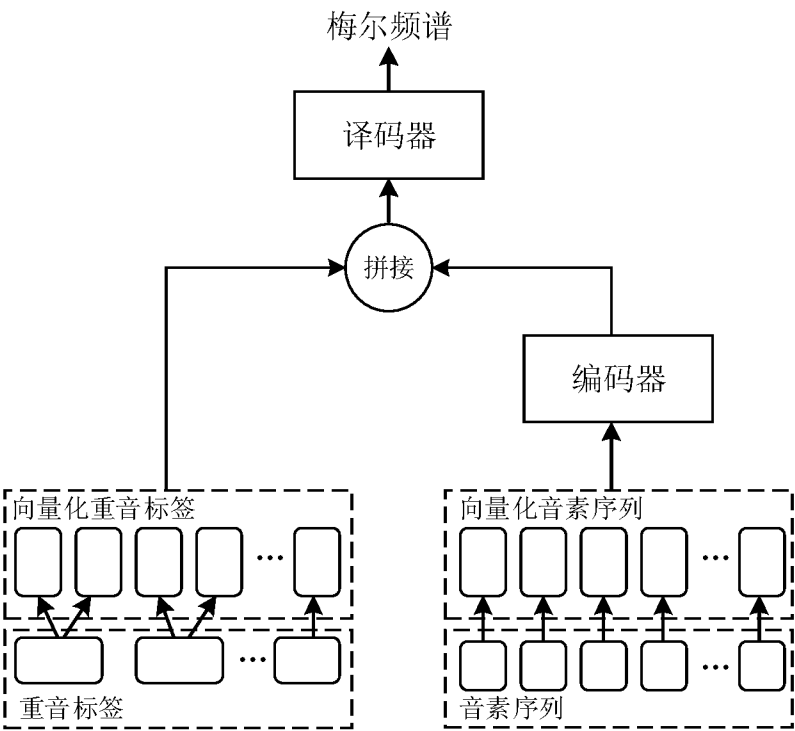


图 3

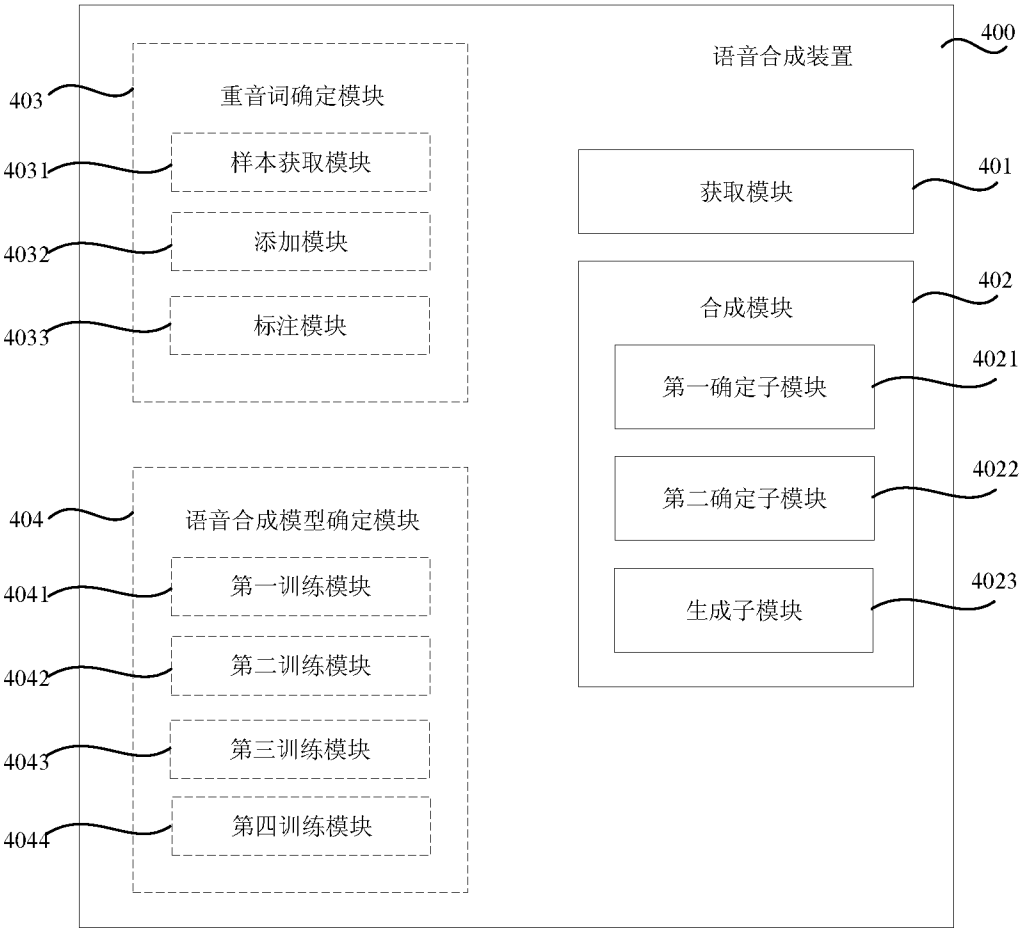


图 4

500

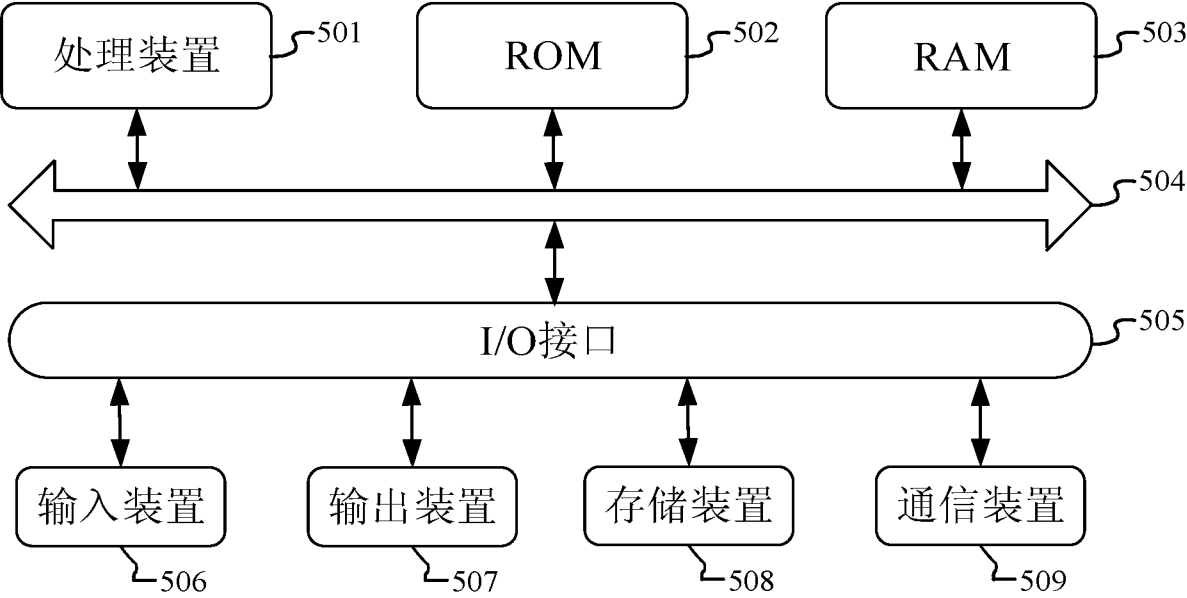


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/126394

A. CLASSIFICATION OF SUBJECT MATTER

G10L 13/10(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L13;G10L15

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

DWPI; WPABS; CNABS; CNTXT; CNKI; 百度学术: 语音合成, 文语, 重音, 重读, 韵律, 标记, 标志, 标签, 向量, 拼接, 组合, 训练, 合成模型, 深度学习, speech synthesis, tts, emphasis, stress, prosody, label, tab, vector, deep learn

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112331176 A (BEIJING YOUZHUJU NETWORK TECHNOLOGY CO. LTD.) 05 February 2021 (2021-02-05) claims 1-10	1-25
X	CN 110299131 A (SUZHOU QDREAMER NETWORK SCIENCE AND TECH CO., LTD.) 01 October 2019 (2019-10-01) description, paragraphs [0029]-[0054], and figures 1-2	1-7, 11-18, 22-25
X	CN 111667816 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 15 September 2020 (2020-09-15) description, paragraphs [0046]-[0078]	1-7, 11-18, 22-25
Y	CN 110299131 A (SUZHOU QDREAMER NETWORK SCIENCE AND TECH CO., LTD.) 01 October 2019 (2019-10-01) description, paragraphs [0029]-[0054], and figures 1-2	8-10, 19-21
Y	CN 111667816 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 15 September 2020 (2020-09-15) description, paragraphs [0046]-[0078]	8-10, 19-21



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

24 January 2022

Date of mailing of the international search report

30 January 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/126394**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 109087627 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 25 December 2018 (2018-12-25) paragraphs [0036]-[0048]	8-10, 19-21
A	CN 109686359 A (NUBIA TECHNOLOGY CO., LTD.) 26 April 2019 (2019-04-26) entire document	1-25
A	JP 2018169434 A (FUJITSU LIMITED) 01 November 2018 (2018-11-01) entire document	1-25
A	CN 108510975 A (BAIDU USA L.L.C.) 07 September 2018 (2018-09-07) entire document	1-25

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/126394

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	112331176	A	05 February 2021	None			
CN	110299131	A	01 October 2019	None			
CN	111667816	A	15 September 2020	None			
CN	109087627	A	25 December 2018	None			
CN	109686359	A	26 April 2019	None			
JP	2018169434	A	01 November 2018	JP	6790959	B2	25 November 2020
CN	108510975	A	07 September 2018	US	2018247636	A1	30 August 2018
				US	10872598	B2	22 December 2020
				US	2021027762	A1	28 January 2021

A. 主题的分类 G10L 13/10 (2013.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																										
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G10L13; G10L15 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) DWPI; WPABS; CNABS; CNTXT; CNKI; 百度学术: 语音合成, 文语, 重音, 重读, 韵律, 标记, 标志, 标签, 向量, 拼接, 组合, 训练, 合成模型, 深度学习, speech synthesis, tts, emphasis, stress, prosody, label, tab, vector, deep learn																										
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 112331176 A (北京有竹居网络技术有限公司) 2021年2月5日 (2021 - 02 - 05) 权利要求1-10</td> <td>1-25</td> </tr> <tr> <td>X</td> <td>CN 110299131 A (苏州奇梦者网络科技有限公司) 2019年10月1日 (2019 - 10 - 01) 说明书第[0029]-[0054]段, 附图1-2</td> <td>1-7, 11-18, 22-25</td> </tr> <tr> <td>X</td> <td>CN 111667816 A (北京百度网讯科技有限公司) 2020年9月15日 (2020 - 09 - 15) 说明书第[0046]-[0078]段</td> <td>1-7, 11-18, 22-25</td> </tr> <tr> <td>Y</td> <td>CN 110299131 A (苏州奇梦者网络科技有限公司) 2019年10月1日 (2019 - 10 - 01) 说明书第[0029]-[0054]段, 附图1-2</td> <td>8-10, 19-21</td> </tr> <tr> <td>Y</td> <td>CN 111667816 A (北京百度网讯科技有限公司) 2020年9月15日 (2020 - 09 - 15) 说明书第[0046]-[0078]段</td> <td>8-10, 19-21</td> </tr> <tr> <td>Y</td> <td>CN 109087627 A (百度在线网络技术北京有限公司) 2018年12月25日 (2018 - 12 - 25) 第[0036]-[0048]段</td> <td>8-10, 19-21</td> </tr> <tr> <td>A</td> <td>CN 109686359 A (努比亚技术有限公司) 2019年4月26日 (2019 - 04 - 26) 全文</td> <td>1-25</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 112331176 A (北京有竹居网络技术有限公司) 2021年2月5日 (2021 - 02 - 05) 权利要求1-10	1-25	X	CN 110299131 A (苏州奇梦者网络科技有限公司) 2019年10月1日 (2019 - 10 - 01) 说明书第[0029]-[0054]段, 附图1-2	1-7, 11-18, 22-25	X	CN 111667816 A (北京百度网讯科技有限公司) 2020年9月15日 (2020 - 09 - 15) 说明书第[0046]-[0078]段	1-7, 11-18, 22-25	Y	CN 110299131 A (苏州奇梦者网络科技有限公司) 2019年10月1日 (2019 - 10 - 01) 说明书第[0029]-[0054]段, 附图1-2	8-10, 19-21	Y	CN 111667816 A (北京百度网讯科技有限公司) 2020年9月15日 (2020 - 09 - 15) 说明书第[0046]-[0078]段	8-10, 19-21	Y	CN 109087627 A (百度在线网络技术北京有限公司) 2018年12月25日 (2018 - 12 - 25) 第[0036]-[0048]段	8-10, 19-21	A	CN 109686359 A (努比亚技术有限公司) 2019年4月26日 (2019 - 04 - 26) 全文	1-25
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																								
PX	CN 112331176 A (北京有竹居网络技术有限公司) 2021年2月5日 (2021 - 02 - 05) 权利要求1-10	1-25																								
X	CN 110299131 A (苏州奇梦者网络科技有限公司) 2019年10月1日 (2019 - 10 - 01) 说明书第[0029]-[0054]段, 附图1-2	1-7, 11-18, 22-25																								
X	CN 111667816 A (北京百度网讯科技有限公司) 2020年9月15日 (2020 - 09 - 15) 说明书第[0046]-[0078]段	1-7, 11-18, 22-25																								
Y	CN 110299131 A (苏州奇梦者网络科技有限公司) 2019年10月1日 (2019 - 10 - 01) 说明书第[0029]-[0054]段, 附图1-2	8-10, 19-21																								
Y	CN 111667816 A (北京百度网讯科技有限公司) 2020年9月15日 (2020 - 09 - 15) 说明书第[0046]-[0078]段	8-10, 19-21																								
Y	CN 109087627 A (百度在线网络技术北京有限公司) 2018年12月25日 (2018 - 12 - 25) 第[0036]-[0048]段	8-10, 19-21																								
A	CN 109686359 A (努比亚技术有限公司) 2019年4月26日 (2019 - 04 - 26) 全文	1-25																								
☒ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																										
<table border="0"> <tr> <td> * 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 </td> <td> “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件 </td> </tr> </table>			* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																						
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																									
国际检索实际完成的日期 2022年1月24日		国际检索报告邮寄日期 2022年1月30日																								
ISA/CN的名称和邮寄地址 中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 游晓梅 电话号码 (86-10)62089539																								

C. 相关文件		
类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	JP 2018169434 A (FUJITSU LTD.) 2018年11月1日 (2018 - 11 - 01) 全文	1-25
A	CN 108510975 A (百度美国有限责任公司) 2018年9月7日 (2018 - 09 - 07) 全文	1-25

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/126394

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	112331176	A	2021年2月5日	无			
CN	110299131	A	2019年10月1日	无			
CN	111667816	A	2020年9月15日	无			
CN	109087627	A	2018年12月25日	无			
CN	109686359	A	2019年4月26日	无			
JP	2018169434	A	2018年11月1日	JP	6790959	B2	2020年11月25日
CN	108510975	A	2018年9月7日	US	2018247636	A1	2018年8月30日
				US	10872598	B2	2020年12月22日
				US	2021027762	A1	2021年1月28日