

(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
2020 年 11 月 26 日 (26.11.2020)



(10) 国际公布号
WO 2020/232997 A1

(51) 国际专利分类号:
G10L 13/08 (2013.01)

(21) 国际申请号: PCT/CN2019/117254

(22) 国际申请日: 2019 年 11 月 11 日 (11.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910420154.9 2019 年 5 月 20 日 (20.05.2019) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518057 (CN)。

(72) 发明人: 王健宗(WANG, Jianzong); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518057 (CN)。 孙奥兰(SUN, Aolan); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518057 (CN)。 彭话易(PENG, Huayi); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518057 (CN)。 程宁(CHENG, Ning); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518057 (CN)。

(74) 代理人: 深圳市力道知识产权代理有限公司(普通合伙)(SHENZHEN LIDAO INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区

(54) Title: SPEECH SYNTHESIS METHOD AND APPARATUS, AND DEVICE AND COMPUTER-READABLE STORAGE MEDIUM

(54) 发明名称: 语音合成方法、装置、设备及计算机可读存储介质

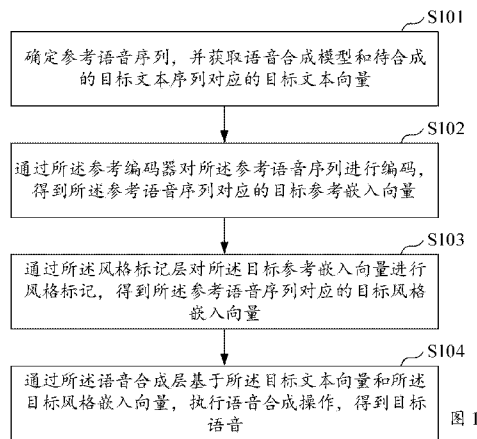


图 1

- S101 Determine a reference speech sequence, and acquire a speech synthesis model and a target text vector corresponding to a target text sequence to be synthesized
- S102 Encode the reference speech sequence by means of a reference encoder to obtain a target reference embedded vector corresponding to the reference speech sequence
- S103 Perform, by means of a style marking layer, style marking on the target reference embedded vector to obtain a target style embedded vector corresponding to the reference speech sequence
- S104 By means of a speech synthesis layer and on the basis of the target text vector and the target style embedded vector, execute a speech synthesis operation to obtain target speech

(57) Abstract: A speech synthesis method and apparatus, and a device and a computer-readable storage medium. The method comprises: determining a reference speech sequence, and acquiring a speech synthesis model and a target text vector corresponding to a target text sequence to be synthesized (S101); encoding the reference speech sequence by means of a reference encoder to obtain a target reference embedded vector corresponding to the reference speech sequence (S102); performing, by means of a style marking layer, style marking on the target reference embedded vector to obtain a target style embedded vector corresponding to the reference speech

园岭街道八卦一路50号鹏基商务时空大厦
1717-1719, Guangdong 518000 (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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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))。

sequence (S103); and by means of a speech synthesis layer and on the basis of the target text vector and the target style embedded vector, executing a speech synthesis operation to obtain target speech (S104). By means of the method, speech synthesized by a speech rhythm expressed by a target style embedded vector can be obtained, thereby effectively improving the expressive power and accuracy of synthetic speech.

(57) 摘要: 一种语音合成方法、装置、设备及计算机可读存储介质, 该方法包括: 确定参考语音序列, 并获取语音合成模型和待合成的目标文本序列对应的目标文本向量(S101); 通过参考编码器对参考语音序列进行编码, 得到参考语音序列对应的目标参考嵌入向量(S102); 通过风格标记层对目标参考嵌入向量进行风格标记, 得到参考语音序列对应的目标风格嵌入向量(S103); 通过语音合成层基于目标文本向量和目标风格嵌入向量, 执行语音合成操作, 得到目标语音(S104)。该方法能够得到目标风格嵌入向量表达的语音韵律合成语音, 有效提高合成语音的表达力和精确性。

语音合成方法、装置、设备及计算机可读存储介质

本申请要求于2019年5月20日提交中国专利局、申请号为201910420154.9、
5 发明名称为“语音合成方法、装置、设备及计算机可读存储介质”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

本申请涉及语音合成的技术领域，尤其涉及一种语音合成方法、装置、
10 设备及计算机可读存储介质。

背景技术

随着 TTS (Text To Speech, 从文本到语音) 系统的快速发展, 使用 TTS
系统的场景也越来越多, 如语音书的朗读、新闻阅读和交谈助手等, 其中,
15 神经网络模型具备合成有表达性效果语音的性能, 但为了合成更像真人的语
音, 神经网络模型必须学习韵律, 而韵律是语音中一组现象的组合, 例如副
语言信息、腔调、重音和风格等。

风格包含了丰富的信息, 如意图、情绪和影响说话者的腔调和语音流等,
因此很难精准的定义语音的风格, 而目前的 TTS 系统以及端到端的语音合成
20 系统, 仅能学习一个输入数据的平均韵律分布, 没有风格的控制, 对于较长
的文本句子无法合成具备精准表达力的语音。因此, 如何提高合成语音的表
达力精确性是目前亟待解决的问题。

发明内容

25 本申请的主要目的在于提供一种语音合成方法、装置、设备及计算机可
读存储介质, 旨在提高合成语音的表达力精确性。

第一方面, 本申请提供一种语音合成方法, 所述语音合成方法包括以下
步骤:

确定参考语音序列, 并获取语音合成模型和待合成的目标文本序列对应
30 的目标文本向量, 其中, 所述语音合成模型至少包括参考编码器、风格标记
层和语音合成层;

通过所述参考编码器对所述参考语音序列进行编码, 得到所述参考语音
序列对应的目标参考嵌入向量;

通过所述风格标记层对所述目标参考嵌入向量进行风格标记, 得到所述

参考语音序列对应的目标风格嵌入向量；

通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音。

第二方面，本申请还提供一种语音合成装置，所述语音合成装置包括：

5 确定模块，用于确定参考语音序列；

获取模块，用于获取语音合成模型和待合成的目标文本序列对应的目标文本向量，其中，所述语音合成模型至少包括参考编码器、风格标记层和语音合成层；

10 编码模块，用于通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量；

风格标记模块，用于通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量；

语音合成模块，用于通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音。

15 第三方面，本申请还提供一种计算机设备，所述计算机设备包括处理器、存储器、以及存储在所述存储器上并可被所述处理器执行的计算机程序，其中所述计算机程序被所述处理器执行时，实现如上述的语音合成方法的步骤。

第四方面，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有计算机程序，其中所述计算机程序被处理器执行时，实现如
20 上述的语音合成方法的步骤。

本申请提供一种语音合成方法、装置、设备及计算机可读存储介质，本申请通过语音合成模型中的参考编码器和风格标记层对参考语音序列进行处理，可以得到目标风格嵌入向量，即语音韵律的向量表达，然后通过语音合成模型中的语音合成层，基于目标风格嵌入向量和目标文本向量，共同合成
25 语音，可以得到按照该目标风格嵌入向量表达的语音韵律合成语音，可以有效的提高合成语音的表达力精确性。

附图说明

30 为了更清楚地说明本申请实施例技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图1为本申请实施例提供的一种语音合成方法的流程示意图；

图2为本申请实施例中的语音合成模型的一层级示意图；

35 图3为本申请实施例提供的另一种语音合成方法的流程示意图；

图4为本申请实施例提供的一种语音合成装置的示意性框图；

图5为本申请实施例提供的另一种语音合成装置的示意性框图；

图6为本申请一实施例涉及的计算机设备的结构示意图。

5 本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

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清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是
10 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得的所有其他实施例，都属于本申请保护的范围。

附图中所示的流程图仅是示例说明，不是必须包括所有的内容和操作/步骤，也不是必须按所描述的顺序执行。例如，有的操作/步骤还可以分解、组合或部分合并，因此实际执行的顺序有可能根据实际情况改变。

15 本申请实施例提供一种语音合成方法、装置、计算机设备及计算机可读存储介质。其中，该语音合成方法可应用于服务器中，该服务器可以为单台的服务器，也可以为服务器集群。

下面结合附图，对本申请的一些实施方式作详细说明。在不冲突的情况下，下述的实施例及实施例中的特征可以相互组合。

20 请参照图 1，图 1 为本申请的实施例提供的一种语音合成方法的流程示意图。

如图 1 所示，该语音合成方法，用于提高合成语音的表达力精确性，其中该语音合成方法包括步骤 S101 至步骤 S104。

25 步骤 S101、确定参考语音序列，并获取语音合成模型和待合成的目标文本序列对应的目标文本向量，其中，所述语音合成模型至少包括参考编码器、风格标记层和语音合成层。

30 在需要合成具有精确表达力的语音时，需要先确定参与语音合成的参考语音序列，然后按照参考语音序列的语音韵律来合成文本对应的语音，为此可以预存至少一个语音序列，在需要合成语音时，将预存的至少一个语音序列确定为参考语音序列。其中，语音序列可以是音频信号，该音频信号可以是变长的或者是固定长度的，该语音序列可由音素组成，音素是构成音节的最小单位或最小的语音片段，是从音质的角度划分出来的最小的线性的语音单位。

35 在一实施例中，为了灵活的改变合成语音的韵律，可以通过选择的方式确定参考语音序列，具体为当监测到语音合成请求时，从该语音合成请求中

获取语音风格编号，并获取该语音风格标号对应的语音序列，然后将该语音序列确定为参考语音序列。其中，服务器中存储有语音风格编号与语音序列的映射关系表，通过查询该映射关系表即可得到每个语音风格编号对应的语音序列。

5 在确定参考语音序列的同时，获取语音合成模型和参与语音合成的目标文本序列对应的目标文本向量，其中，语音合成模型至少包括参考编码器、风格标记层和语音合成层，该语音合成模型是通过训练得到的，且训练过程为：确定待训练的语音合成模型，包括参考编码器、风格标记层和语音合成层，并将作为样本的语音序列输入该参考编码器，得到参考嵌入向量，然后
10 将该参考嵌入向量输入该风格标记层，得到风格嵌入向量，并将作为样本的文本向量与该风格嵌入向量输入至该语音合成层，从而训练语音合成模型，通过不断迭代训练直到语音合成模型收敛。其中，该语音序列所表达的内容与该文本向量表达的内容相同。

步骤 S102、通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量。
15

在确定参考语音序列之后，该服务器通过该语音合成模型中的参考编码器对该参考语音序列进行编码，得到该参考语音序列对应的目标参考嵌入向量。在一实施例中，该参考编码器由卷积神经网络（Convolutional Neural Networks, CNN）层和循环神经网络（Recurrent Neural Network, RNN）层组成，且该卷积神经网络层由多层二维卷积层组成，该循环神经网络层由一个
20 RNN 组成，其中，该二维卷积层的核可选为 3×3 ，步长可选为 2×2 ，例如，如果该 CNN 层为 6 层二维卷积层，则可以为这六层的二维卷积层依次设置 32、32、64、64、128 和 128 的输出通道。

具体地，通过该参考编码器中的卷积神经网络层对该参考语音序列进行处理，得到该参考语音序列对应的三维张量，即从该参考语音序列中提取音频特征，并依次通过该卷积神经网络层中的每个二维卷积层处理该音频特征，得到张量，并将该张量变换为三维张量，但保持输出的时间复杂度；然后通过该参考编码器中的循环神经网络层对该三维张量进行处理，得到该参考语音序列对应的目标参考嵌入向量。
25

步骤 S103、通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量。
30

在得到目标参考嵌入向量之后，该服务器通过该语音合成模型的风格标记层对该目标参考嵌入向量进行风格标记，得到该参考语音序列对应的目标风格嵌入向量。其中，该风格标记层中存储有每个风格标记对参考语音序列贡献的风格权重，且每个风格标记对参考语音序列贡献的风格权重之和为 1，
35

需要说明的是，每个风格标记对参考语音序列贡献的风格权重是在训练语音合成模型收敛时得到的。

具体地，从该风格标记层获取每个风格标记对该参考语音序列贡献的风格权重，并用每个风格标记对该参考语音序列贡献的风格权重乘以该目标参考嵌入向量，得到每个风格标记的风格嵌入向量，然后将每个风格标记的风格嵌入向量之和确定为该参考语音序列对应的目标风格嵌入向量。需要说明的是，该每个风格标记的风格权重共同构成一个权重组，表示一种语音韵律，可以控制参考语音序列的语音合成，使得参考语音序列按照该权重组表达的语音韵律播放。

步骤 S104、通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音。

在得到目标风格嵌入向量之后，该服务器通过语音合成模型中的语音合成层基于该目标文本向量和该目标风格嵌入向量，执行语音合成操作，得到目标语音。请参照图 2，图 2 为本申请实施例中语音合成模型的一层级示意图，如图 2 所示，该语音合成模型包括参考编码器、风格标记层和语音合成层，参考编码器的输入为参考语音序列，语音合成层的输入为目标风格嵌入向量和目标文本向量，语音合成层的输出为目标语音。

在一实施例中，该语音合成层包括向量拼接子层和语音合成子层，可选地，该语音合成子层由端对端的语音合成模型，即 Tacotron 模型组成，该 Tacotron 模型依次由编码器-注意力机制-解码器组成。具体地，通过该语音合成层中的向量拼接子层对该目标文本向量和该目标风格嵌入向量进行拼接，得到该目标文本序列的拼接向量，并通过该语音合成层中的语音合成子层基于该拼接向量，执行语音合成操作，得到该目标文本序列对应的目标语音。

在一实施例中，为提高语音合成速度，在语音合成层中加入状态标识符，当状态标识符为合成结束状态对应的标识符时，结束语音合成，具体地，将目标文本向量和目标风格嵌入向量输入语音合成层，执行语音合成操作，然后在执行语音合成操作的过程中，实时监测该语音合成层的状态标识符是否为合成结束状态对应的标识符，当监测该语音合成层的状态标识符为合成结束状态对应的标识符时，停止执行语音合成操作，并输出该目标文本序列对应的目标语音。

上述实施例提供的语音合成方法，通过语音合成模型中的参考编码器和风格标记层对参考语音序列进行处理，可以得到目标风格嵌入向量，即语音韵律的向量表达，然后通过语音合成模型中的语音合成层，基于目标风格嵌入向量和目标文本向量，共同合成语音，可以得到按照该目标风格嵌入向量表达的语音韵律合成语音，可以有效的提高合成语音的表达力精确性。

请参照图 3, 图 3 为本申请实施例提供的另一种语音合成方法的流程示意图。

如图 3 所示, 该语音合成方法包括步骤 S201 至 204。

步骤 S201、确定参考语音序列, 并获取语音合成模型和待合成的目标文本序列对应的目标文本向量, 其中, 所述语音合成模型至少包括参考编码器、风格标记层和语音合成层。

在需要合成具有精确表达力的语音时, 需要先确定参与语音合成的参考语音序列, 然后按照参考语音序列的语音韵律来合成文本对应的语音, 为此可以预存至少一个语音序列, 在需要合成语音时, 将预存的至少一个语音序列确定为参考语音序列。其中, 语音序列可以是音频信号, 该音频信号可以是变长的或者是固定长度的, 该语音序列可由音素组成, 音素是构成音节的最小单位或最小的语音片段, 是从音质的角度划分出来的最小的线性的语音单位。其中, 语音合成模型至少包括参考编码器、风格标记层和语音合成层, 该语音合成模型是通过训练得到的。

步骤 S202、通过所述参考编码器对所述参考语音序列进行编码, 得到所述参考语音序列对应的目标参考嵌入向量。

在确定参考语音序列之后, 该服务器通过该语音合成模型中的参考编码器对该参考语音序列进行编码, 得到该参考语音序列对应的目标参考嵌入向量。具体地, 通过该参考编码器中的卷积神经网络层对该参考语音序列进行处理, 得到该参考语音序列对应的三维张量, 即从该参考语音序列中提取音频特征, 并依次通过该卷积神经网络层中的每个二维卷积层处理该音频特征, 得到张量, 并将该张量变换为三维张量, 但保持输出的时间复杂度; 然后通过该参考编码器中的循环神经网络层对该三维张量进行处理, 得到该参考语音序列对应的目标参考嵌入向量。

步骤 S203、通过所述风格标记层对所述目标参考嵌入向量进行风格标记, 得到所述参考语音序列对应的目标风格嵌入向量。

在得到目标参考嵌入向量之后, 该服务器通过该语音合成模型的风格标记层对该目标参考嵌入向量进行风格标记, 得到该参考语音序列对应的目标风格嵌入向量。目标风格嵌入向量表示一种参考语音序列的语音韵律, 为准确的确定参考语音序列的语音韵律表示, 具体地, 参照图 3, 步骤 S203 包括子步骤 S2031 至 S2033。

子步骤 S2031、通过所述风格标记层中的注意力机制, 计算每个风格标记对应的预设向量与所述目标参考嵌入向量之间的相似度。

其中, 在训练语音合成模型的过程中, 风格标记层中的注意力机制也一起训练, 在语音合成模型收敛时, 风格标记层中的注意力机制也相应的收敛,

同时获取收敛时每个风格标记对应的参考嵌入向量，然后将每个风格标记对应的参考嵌入向量作为预设向量固化在风格标记层中。

在得到目标参考嵌入向量之后，通过风格标记层中的注意力机制，计算每个风格标记对应的预设向量与该目标参考嵌入向量之间的相似度。其中，
5 该注意力机制可以是多头注意力机制，或者是基于内容的注意力机制，或者是基于位置的注意力机制，或者是基于几种注意力机制的结合。

子步骤 S2032、根据每个风格标记对应的预设向量与所述目标参考嵌入向量之间的相似度，确定每个风格标记对所述参考语音序列贡献的风格权重。

在确定每个风格标记对应的预设向量与所述目标参考嵌入向量之间的相似
10 度之后，根据每个风格标记对应的预设向量与该目标参考嵌入向量之间的相似度，确定每个风格标记对参考语音序列贡献的风格权重，即累加每个风格标记对应的预设向量与该目标参考嵌入向量之间的相似度，得到总相似度，并计算每个风格标记对应的预设向量与该目标参考嵌入向量之间的相似度占该总相似度的比率，然后将每个风格标记对应的预设向量与该目标参考嵌入
15 向量之间的相似度占该总相似度的比率确定为每个风格标记对参考语音序列贡献的风格权重。

例如，风格标记的数量为 5，且每个风格标记对应的预设向量与该目标参考嵌入向量之间的相似度分别为 0.6、0.3、0.4、0.4 和 0.3，则总相似度为 2，每个风格标记的该相似度与总相似度之比分别为 0.3、0.15、0.2、0.2 和 0.15，
20 则 每个风格标记对参考语音序列贡献的风格权重分别为 0.3、0.15、0.2、0.2 和 0.15。

子步骤 S2033、根据每个风格标记对所述参考语音序列贡献的风格权重，确定所述参考语音序列对应的目标风格嵌入向量。

在确定每个风格标记对该参考语音序列贡献的风格权重之后，用每个风格
25 标记对该参考语音序列贡献的风格权重乘以该目标参考嵌入向量，得到每个风格标记的风格嵌入向量，然后累加每个风格标记的风格嵌入向量，得到该参考语音序列对应的目标风格嵌入向量。

步骤 S204、通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音。

在得到目标风格嵌入向量之后，该服务器通过语音合成模型中的语音合成
30 层基于该目标文本向量和该目标风格嵌入向量，执行语音合成操作，得到目标语音。具体地，通过该语音合成层中的向量拼接子层对该目标文本向量和该目标风格嵌入向量进行拼接，得到该目标文本序列的拼接向量，并通过该语音合成层中的语音合成子层基于该拼接向量，执行语音合成操作，得到
35 该目标文本序列对应的目标语音。

上述实施例提供的语音合成方法，通过风格标记层中的注意力机制，可以准确计算出每个风格标记对参考语音序列贡献的风格权重，从而可以准确的确定参考语音序列的语音韵律表示，同时通过语音合成模型中的语音合成层，基于目标风格嵌入向量和目标文本向量，共同合成为语音，可以得到按照该目标风格嵌入向量表达的语音韵律合成为语音，可以有效的提高合成为语音的表达力精确性。

请参照图4，图4为本申请实施例提供的一种语音合成装置的示意性框图。

如图4所示，该语音合成装置300，包括：确定模块301、获取模块302、编码模块303、风格标记模块304和语音合成模块305。

确定模块301，用于确定参考语音序列。

在一实施例中，所述确定模块301，还用于当监测到语音合成请求时，从所述语音合成请求中获取语音风格编号；获取所述语音风格编号对应的语音序列，并将所述语音序列确定为参考语音序列。

获取模块302，用于获取语音合成模型和待合成的目标文本序列对应的目标文本向量，其中，所述语音合成模型至少包括参考编码器、风格标记层和语音合成。

编码模块303，用于通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量。

在一实施例中，所述编码模块303，还用于通过所述参考编码器中的卷积神经网络层对所述参考语音序列进行处理，得到所述参考语音序列对应的三维张量；通过所述参考编码器中的循环神经网络层对所述三维张量进行处理，得到所述参考语音序列对应的目标参考嵌入向量。

风格标记模块304，用于通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量。

在一实施例中，所述风格标记模块304，还用于从所述风格标记层中获取每个风格标记对所述参考语音序列贡献的风格权重；用每个风格标记对所述参考语音序列贡献的风格权重乘以所述目标参考嵌入向量，得到每个风格标记的风格嵌入向量；累加每个风格标记的风格嵌入向量，得到所述参考语音序列对应的目标风格嵌入向量。

语音合成模块305，用于通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音。

在一实施例中，所述语音合成模块305，还用于通过所述语音合成层中的向量拼接子层对所述目标文本向量和所述目标风格嵌入向量进行拼接，得到所述目标文本序列的拼接向量；通过所述语音合成层中的语音合成子层基于所述拼接向量，执行语音合成操作，得到所述目标文本序列对应的目标语音。

在一实施例中，所述语音合成模块 305，还用于将所述目标文本向量和所述目标风格嵌入向量输入所述语音合成层，执行语音合成操作；在执行语音合成操作的过程中，实时监测所述语音合成层的状态标识符是否为合成结束状态对应的标识符；当监测所述语音合成层的状态标识符为合成结束状态对应的标识符时，停止执行语音合成操作，并输出所述目标文本序列对应的目标语音。

请参照图 5，图 5 为本申请实施例提供的另一种语音合成装置的示意性框图。

如图 5 所示，该语音合成装置 400，包括：确定模块 401、获取模块 402、编码模块 403、风格标记模块 404 和语音合成模块 405。

确定模块 401，用于确定参考语音序列。

获取模块 402，用于获取语音合成模型和待合成的目标文本序列对应的目标文本向量，其中，所述语音合成模型至少包括参考编码器、风格标记层和语音合成。

编码模块 403，用于通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量。

风格标记模块 404，用于通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量。

在一实施例中，如图 5 所示，所述风格标记模块 404 包括：

计算子模块 4041，用于通过所述风格标记层中的注意力机制，计算每个风格标记对应的预设向量与所述目标参考嵌入向量之间的相似度。

权重确定子模块 4042，用于根据每个风格标记对应的预设向量与所述目标参考嵌入向量之间的相似度，确定每个风格标记对所述参考语音序列贡献的风格权重。

向量确定子模块 4043，用于根据每个风格标记对所述参考语音序列贡献的风格权重，确定所述参考语音序列对应的目标风格嵌入向量。

语音合成模块 405，用于通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音。

需要说明的是，所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，上述描述的装置和各模块及单元的具体工作过程，可以参考前述语音合成方法实施例中的对应过程，在此不再赘述。

上述实施例提供的装置可以实现为一种计算机程序的形式，该计算机程序可以在如图 6 所示的计算机设备上运行。

请参阅图 6，图 6 为本申请实施例提供的一种计算机设备的结构示意图。该计算机设备可以为服务器。

如图 6 所示, 该计算机设备包括通过系统总线连接的处理器、存储器和网络接口, 其中, 存储器可以包括非易失性存储介质和内存存储器。

非易失性存储介质可存储操作系统和计算机程序。该计算机程序包括程序指令, 该程序指令被执行时, 可使得处理器执行任意一种语音合成方法。

5 处理器用于提供计算和控制能力, 支撑整个计算机设备的运行。

内存存储器为非易失性存储介质中的计算机程序的运行提供环境, 该计算机程序被处理器执行时, 可使得处理器执行任意一种语音合成方法。

该网络接口用于进行网络通信, 如发送分配的任务等。本领域技术人员可以理解, 图 6 中示出的结构, 仅仅是与本申请方案相关的部分结构的框图, 并不构成对本申请方案所应用于其上的计算机设备的限定, 具体的计算机设备可以包括比图中所示更多或更少的部件, 或者组合某些部件, 或者具有不同的部件布置。

应当理解的是, 处理器可以是中央处理单元 (Central Processing Unit, CPU), 该处理器还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现场可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。其中, 通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

其中, 在一个实施例中, 所述处理器用于运行存储在存储器中的计算机程序, 以实现如下步骤:

确定参考语音序列, 并获取语音合成模型和待合成的目标文本序列对应的目标文本向量, 其中, 所述语音合成模型至少包括参考编码器、风格标记层和语音合成层;

通过所述参考编码器对所述参考语音序列进行编码, 得到所述参考语音序列对应的目标参考嵌入向量;

通过所述风格标记层对所述目标参考嵌入向量进行风格标记, 得到所述参考语音序列对应的目标风格嵌入向量;

通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量, 执行语音合成操作, 得到目标语音。

30 在一个实施例中, 所述处理器在实现通过所述参考编码器对所述参考语音序列进行编码, 得到所述参考语音序列对应的目标参考嵌入向量时, 用于实现:

通过所述参考编码器中的卷积神经网络层对所述参考语音序列进行处理, 得到所述参考语音序列对应的三维张量;

35 通过所述参考编码器中的循环神经网络层对所述三维张量进行处理, 得

到所述参考语音序列对应的目标参考嵌入向量。

在一个实施例中，所述处理器在实现通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量时，用于实现：

- 5 从所述风格标记层中获取每个风格标记对所述参考语音序列贡献的风格权重；

用每个风格标记对所述参考语音序列贡献的风格权重乘以所述目标参考嵌入向量，得到每个风格标记的风格嵌入向量；

- 10 累加每个风格标记的风格嵌入向量，得到所述参考语音序列对应的目标风格嵌入向量。

在一个实施例中，所述处理器在实现通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音时，用于实现：

- 15 通过所述语音合成层中的向量拼接子层对所述目标文本向量和所述目标风格嵌入向量进行拼接，得到所述目标文本序列的拼接向量；

通过所述语音合成层中的语音合成子层基于所述拼接向量，执行语音合成操作，得到所述目标文本序列对应的目标语音。

- 20 在一个实施例中，所述处理器在实现通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音时，用于实现：

将所述目标文本向量和所述目标风格嵌入向量输入所述语音合成层，执行语音合成操作；

在执行语音合成操作的过程中，实时监测所述语音合成层的状态标识符是否为合成结束状态对应的标识符；

- 25 当监测所述语音合成层的状态标识符为合成结束状态对应的标识符时，停止执行语音合成操作，并输出所述目标文本序列对应的目标语音。

在一个实施例中，所述处理器在实现确定参考语音序列时，用于实现：

当监测到语音合成请求时，从所述语音合成请求中获取语音风格编号；

- 30 获取所述语音风格编号对应的语音序列，并将所述语音序列确定为参考语音序列。

其中，在另一实施例中，所述处理器用于运行存储在存储器中的计算机程序，以实现通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量的步骤包括：

- 35 通过所述风格标记层中的注意力机制，计算每个风格标记对应的预设向量与所述目标参考嵌入向量之间的相似度；

根据每个风格标记对应的预设向量与所述目标参考嵌入向量之间的相似度，确定每个风格标记对所述参考语音序列贡献的风格权重；

根据每个风格标记对所述参考语音序列贡献的风格权重，确定所述参考语音序列对应的目标风格嵌入向量。

5 本申请实施例还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有计算机程序，所述计算机程序中包括程序指令，所述程序指令被执行时所实现的方法可参照本申请语音合成方法的各个实施例。

10 其中，所述计算机可读存储介质可以是前述实施例所述的计算机设备的内部存储单元，例如所述计算机设备的硬盘或内存。所述计算机可读存储介质也可以是所述计算机设备的外部存储设备，例如所述计算机设备上配备的插接式硬盘，智能存储卡(Smart Media Card, SMC)，安全数字(Secure Digital, SD)卡，闪存卡(Flash Card)等。

15 需要说明的是，在本文中，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、物品或者系统不仅包括那些要素，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、物品或者系统所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者系统中还存在另外的相同要素。

20 上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以权利要求的保护范围为准。

权 利 要 求 书

1、一种语音合成方法，包括：

5 确定参考语音序列，并获取语音合成模型和待合成的目标文本序列对应的目标文本向量，其中，所述语音合成模型至少包括参考编码器、风格标记层和语音合成层；

通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量；

10 通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量；

通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音。

2、如权利要求1所述的语音合成方法，其中，所述通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量的步骤包括：

通过所述参考编码器中的卷积神经网络层对所述参考语音序列进行处理，得到所述参考语音序列对应的三维张量；

通过所述参考编码器中的循环神经网络层对所述三维张量进行处理，得到所述参考语音序列对应的目标参考嵌入向量。

20 3、如权利要求1所述的语音合成方法，其中，所述通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量的步骤包括：

通过所述风格标记层中的注意力机制，计算每个风格标记对应的预设向量与所述目标参考嵌入向量之间的相似度；

25 根据每个风格标记对应的预设向量与所述目标参考嵌入向量之间的相似度，确定每个风格标记对所述参考语音序列贡献的风格权重；

根据每个风格标记对所述参考语音序列贡献的风格权重，确定所述参考语音序列对应的目标风格嵌入向量。

30 4、如权利要求1所述的语音合成方法，其中，所述通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量的步骤包括：

从所述风格标记层中获取每个风格标记对所述参考语音序列贡献的风格权重；

35 用每个风格标记对所述参考语音序列贡献的风格权重乘以所述目标参考嵌入向量，得到每个风格标记的风格嵌入向量；

累加每个风格标记的风格嵌入向量，得到所述参考语音序列对应的目标风格嵌入向量。

5 5、如权利要求 1 所述的语音合成方法，其中，所述通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音的步骤包括：

通过所述语音合成层中的向量拼接子层对所述目标文本向量和所述目标风格嵌入向量进行拼接，得到所述目标文本序列的拼接向量；

通过所述语音合成层中的语音合成子层基于所述拼接向量，执行语音合成操作，得到所述目标文本序列对应的目标语音。

10 6、如权利要求 1-5 中任一项所述的语音合成方法，其中，所述通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音的步骤包括：

将所述目标文本向量和所述目标风格嵌入向量输入所述语音合成层，执行语音合成操作；

15 在执行语音合成操作的过程中，实时监测所述语音合成层的状态标识符是否为合成结束状态对应的标识符；

当监测所述语音合成层的状态标识符为合成结束状态对应的标识符时，停止执行语音合成操作，并输出所述目标文本序列对应的目标语音。

20 7、如权利要求 1-5 中任一项所述的语音合成方法，其中，所述确定参考语音序列的步骤包括：

当监测到语音合成请求时，从所述语音合成请求中获取语音风格编号；

获取所述语音风格编号对应的语音序列，并将所述语音序列确定为参考语音序列。

25 8、一种语音合成装置，其中，所述语音合成装置包括：
确定模块，用于确定参考语音序列；

获取模块，用于获取语音合成模型和待合成的目标文本序列对应的目标文本向量，其中，所述语音合成模型至少包括参考编码器、风格标记层和语音合成层；

30 编码模块，用于通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量；

风格标记模块，用于通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量；

语音合成模块，用于通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音。

35 9、一种计算机设备，其中，所述计算机设备包括处理器、存储器、以及

存储在所述存储器上并可被所述处理器执行的计算机程序，其中所述计算机程序被所述处理器执行时，实现以下步骤：

确定参考语音序列，并获取语音合成模型和待合成的目标文本序列对应的目标文本向量，其中，所述语音合成模型至少包括参考编码器、风格标记层和语音合成层；

通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量；

通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量；

通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音。

10、如权利要求9所述的计算机设备，其中，所述处理器在实现通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量时，用于实现：

通过所述参考编码器中的卷积神经网络层对所述参考语音序列进行处理，得到所述参考语音序列对应的三维张量；

通过所述参考编码器中的循环神经网络层对所述三维张量进行处理，得到所述参考语音序列对应的目标参考嵌入向量。

11、如权利要求9所述的计算机设备，其中，所述处理器在实现通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量时，用于实现：

从所述风格标记层中获取每个风格标记对所述参考语音序列贡献的风格权重；

用每个风格标记对所述参考语音序列贡献的风格权重乘以所述目标参考嵌入向量，得到每个风格标记的风格嵌入向量；

累加每个风格标记的风格嵌入向量，得到所述参考语音序列对应的目标风格嵌入向量。

12、如权利要求9所述的计算机设备，其中，所述处理器在实现通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量时，用于实现：

通过所述语音合成层中的向量拼接子层对所述目标文本向量和所述目标风格嵌入向量进行拼接，得到所述目标文本序列的拼接向量；

通过所述语音合成层中的语音合成子层基于所述拼接向量，执行语音合成操作，得到所述目标文本序列对应的目标语音。

13、如权利要求9所述的计算机设备，其中，所述处理器在实现通过所

述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音时，用于实现：

将所述目标文本向量和所述目标风格嵌入向量输入所述语音合成层，执行语音合成操作；

5 在执行语音合成操作的过程中，实时监测所述语音合成层的状态标识符是否为合成结束状态对应的标识符；

当监测所述语音合成层的状态标识符为合成结束状态对应的标识符时，停止执行语音合成操作，并输出所述目标文本序列对应的目标语音。

14、如权利要求 9-13 中任一项所述的计算机设备，其中，所述处理器在实现确定参考语音序列时，用于实现：

当监测到语音合成请求时，从所述语音合成请求中获取语音风格编号；

获取所述语音风格编号对应的语音序列，并将所述语音序列确定为参考语音序列。

15、一种计算机可读存储介质，其中，所述计算机可读存储介质上存储有计算机程序，其中所述计算机程序被处理器执行时，实现如下步骤：

确定参考语音序列，并获取语音合成模型和待合成的目标文本序列对应的目标文本向量，其中，所述语音合成模型至少包括参考编码器、风格标记层和语音合成层；

20 通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量；

通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量；

通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音。

25 16、如权利要求 15 所述的计算机可读存储介质，其中，所述处理器在实现通过所述参考编码器对所述参考语音序列进行编码，得到所述参考语音序列对应的目标参考嵌入向量时，用于实现：

通过所述参考编码器中的卷积神经网络层对所述参考语音序列进行处理，得到所述参考语音序列对应的三维张量；

30 通过所述参考编码器中的循环神经网络层对所述三维张量进行处理，得到所述参考语音序列对应的目标参考嵌入向量。

17、如权利要求 15 所述的计算机可读存储介质，其中，所述处理器在实现通过所述风格标记层对所述目标参考嵌入向量进行风格标记，得到所述参考语音序列对应的目标风格嵌入向量时，用于实现：

35 从所述风格标记层中获取每个风格标记对所述参考语音序列贡献的风格

权重；

用每个风格标记对所述参考语音序列贡献的风格权重乘以所述目标参考嵌入向量，得到每个风格标记的风格嵌入向量；

5 累加每个风格标记的风格嵌入向量，得到所述参考语音序列对应的目标风格嵌入向量。

18、如权利要求 15 所述的计算机可读存储介质，其中，所述处理器在实现通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音时，用于实现：

10 通过所述语音合成层中的向量拼接子层对所述目标文本向量和所述目标风格嵌入向量进行拼接，得到所述目标文本序列的拼接向量；

通过所述语音合成层中的语音合成子层基于所述拼接向量，执行语音合成操作，得到所述目标文本序列对应的目标语音。

15 19、如权利要求 15 所述的计算机可读存储介质，其中，所述处理器在实现通过所述语音合成层基于所述目标文本向量和所述目标风格嵌入向量，执行语音合成操作，得到目标语音时，用于实现：

将所述目标文本向量和所述目标风格嵌入向量输入所述语音合成层，执行语音合成操作；

在执行语音合成操作的过程中，实时监测所述语音合成层的状态标识符是否为合成结束状态对应的标识符；

20 当监测所述语音合成层的状态标识符为合成结束状态对应的标识符时，停止执行语音合成操作，并输出所述目标文本序列对应的目标语音。

20、如权利要求 15-19 中任一项所述的计算机可读存储介质，其中，所述处理器在实现确定参考语音序列时，用于实现：

当监测到语音合成请求时，从所述语音合成请求中获取语音风格编号；

25 获取所述语音风格编号对应的语音序列，并将所述语音序列确定为参考语音序列。

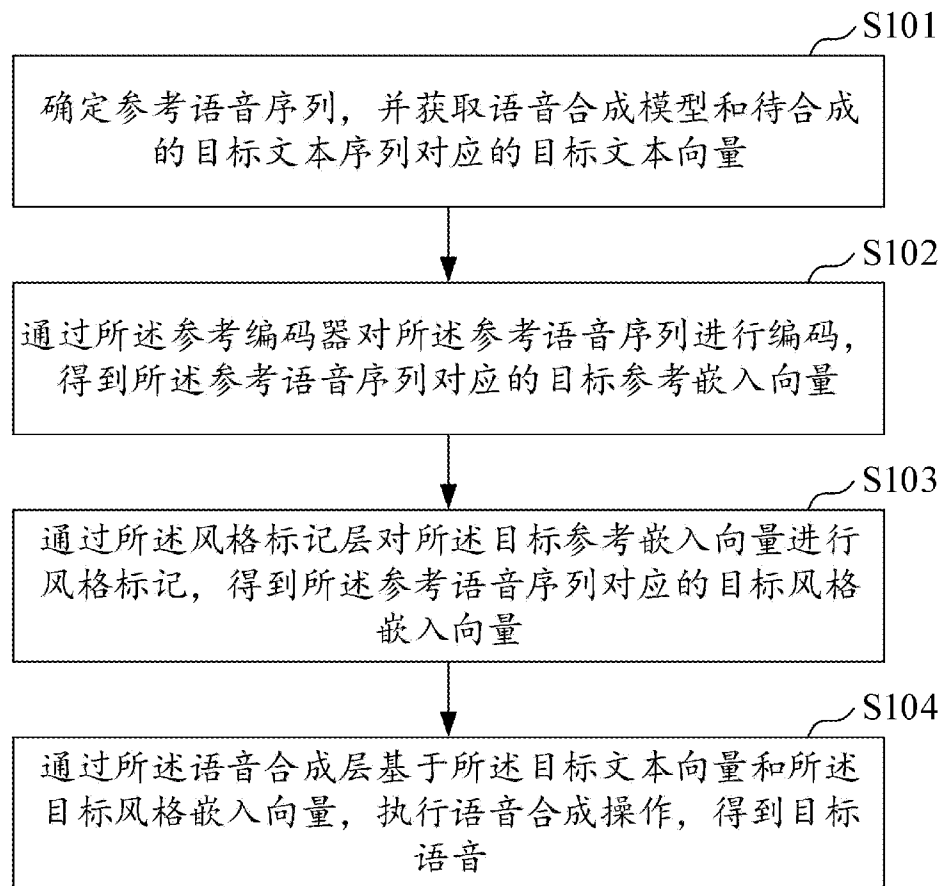


图 1

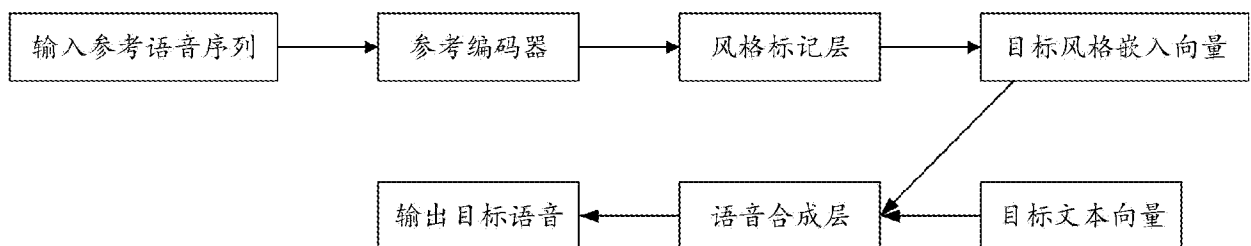


图 2

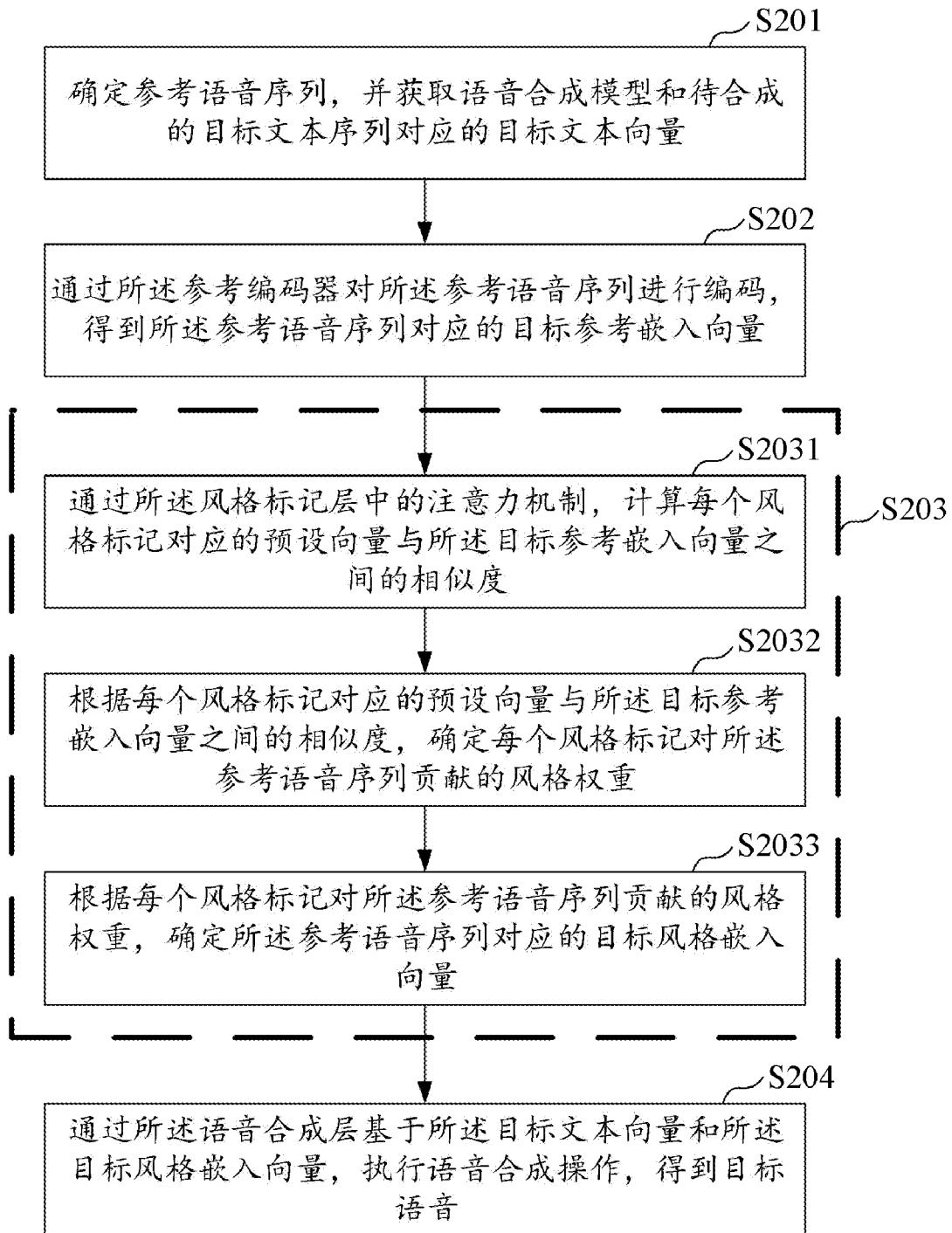


图 3

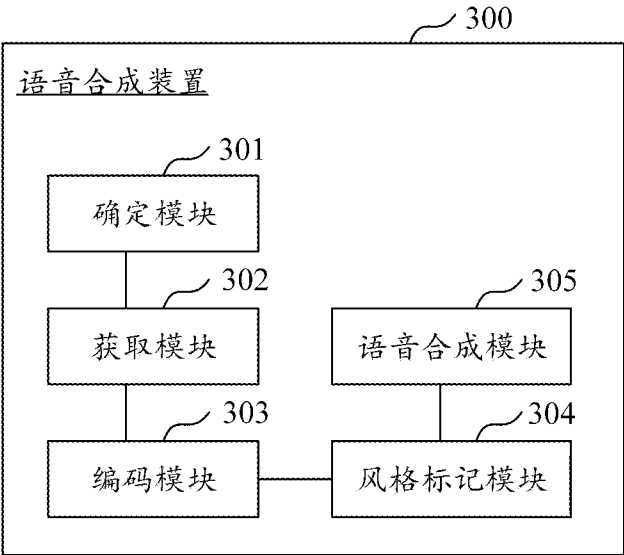


图 4

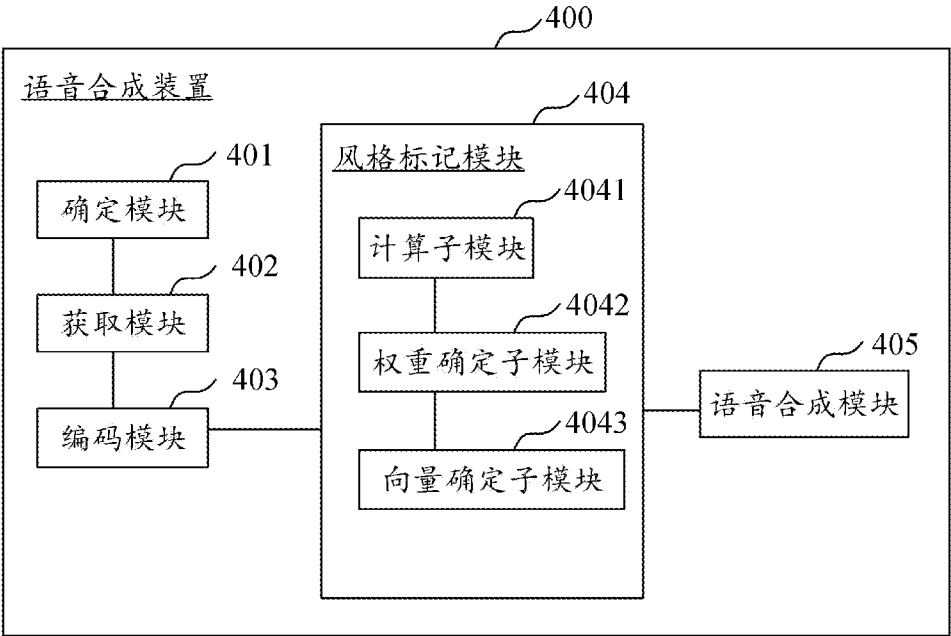


图 5

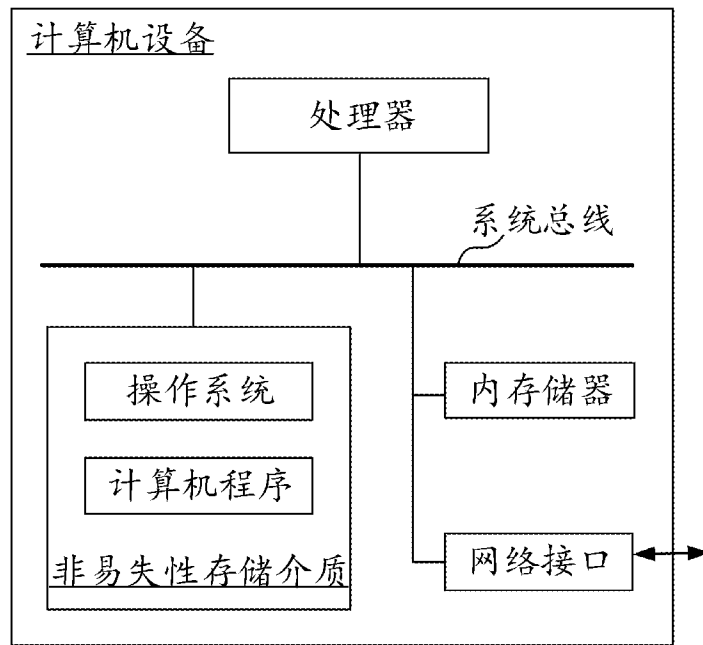


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/117254

A. CLASSIFICATION OF SUBJECT MATTER G10L 13/08(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) G10L 13 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) CNABS; DWPI; 百度学术; 知网: 语音合成, 风格, 韵律, 标记, 嵌入向量, 目标, 迁移, 转换, text to speech, tts, speech synthesis, prosody, style, token, tag, label, embedding, target, transform, convert																				
C. DOCUMENTS CONSIDERED TO BE RELEVANT <table border="1"> <thead> <tr> <th>Category*</th> <th>Citation of document, with indication, where appropriate, of the relevant passages</th> <th>Relevant to claim No.</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 110288973 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 27 September 2019 (2019-09-27) entire document</td> <td>1-20</td> </tr> <tr> <td>X</td> <td>Y. Wang et al. "tyle Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis" <i>arXiv preprint arXiv:1803.09017</i>, 23 March 2018 (2018-03-23), sections 2-4</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 109616127 A (JIANHU YUNFEI DATA TECHNOLOGY CO., LTD.) 12 April 2019 (2019-04-12) entire document</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 109754779 A (CHUMEN WENWEN INFORMATION TECHNOLOGY CO., LTD.) 14 May 2019 (2019-05-14) entire document</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>US 2018336880 A1 (BAIDU USA LLC et al.) 22 November 2018 (2018-11-22) entire document</td> <td>1-20</td> </tr> </tbody> </table>			Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.	PX	CN 110288973 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 27 September 2019 (2019-09-27) entire document	1-20	X	Y. Wang et al. "tyle Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis" <i>arXiv preprint arXiv:1803.09017</i> , 23 March 2018 (2018-03-23), sections 2-4	1-20	A	CN 109616127 A (JIANHU YUNFEI DATA TECHNOLOGY CO., LTD.) 12 April 2019 (2019-04-12) entire document	1-20	A	CN 109754779 A (CHUMEN WENWEN INFORMATION TECHNOLOGY CO., LTD.) 14 May 2019 (2019-05-14) entire document	1-20	A	US 2018336880 A1 (BAIDU USA LLC et al.) 22 November 2018 (2018-11-22) entire document	1-20
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.																		
PX	CN 110288973 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 27 September 2019 (2019-09-27) entire document	1-20																		
X	Y. Wang et al. "tyle Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis" <i>arXiv preprint arXiv:1803.09017</i> , 23 March 2018 (2018-03-23), sections 2-4	1-20																		
A	CN 109616127 A (JIANHU YUNFEI DATA TECHNOLOGY CO., LTD.) 12 April 2019 (2019-04-12) entire document	1-20																		
A	CN 109754779 A (CHUMEN WENWEN INFORMATION TECHNOLOGY CO., LTD.) 14 May 2019 (2019-05-14) entire document	1-20																		
A	US 2018336880 A1 (BAIDU USA LLC et al.) 22 November 2018 (2018-11-22) entire document	1-20																		
☒ 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 12 February 2020		Date of mailing of the international search report 18 February 2020																		
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/CN2019/117254

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Ye Jia et al. "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis" <i>arXiv preprint arXiv: 1806.04558v4</i> , 02 January 2019 (2019-01-02), entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/117254

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	110288973	A	27 September 2019	None	
CN	109616127	A	12 April 2019	None	
CN	109754779	A	14 May 2019	None	
US	2018336880	A1	22 November 2018	None	

国际检索报告

国际申请号

PCT/CN2019/117254

A. 主题的分类 G10L 13/08 (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) G10L 13 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)) CNABS; DWPI; 百度学术; 知网; 语音合成, 风格, 韵律, 标记, 嵌入向量, 目标, 迁移, 转换, text to speech, tts, speech synthesis, prosody, style, token, tag, label, embedding, target, transform, convert																							
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 110288973 A (平安科技深圳有限公司) 2019年 9月 27日 (2019 - 09 - 27) 全文</td> <td>1-20</td> </tr> <tr> <td>X</td> <td>Y. Wang 等. "style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis" arXiv preprint arXiv:1803.09017, 2018年 3月 23日 (2018 - 03 - 23), 第2-4部分</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 109616127 A (建湖云飞数据科技有限公司) 2019年 4月 12日 (2019 - 04 - 12) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 109754779 A (出门问问信息科技有限公司) 2019年 5月 14日 (2019 - 05 - 14) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>US 2018336880 A1 (BAIDU USA LLC等) 2018年 11月 22日 (2018 - 11 - 22) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>Ye Jia 等. "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis" arXiv preprint arXiv: 1806.04558v4, 2019年 1月 2日 (2019 - 01 - 02), 全文</td> <td>1-20</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 110288973 A (平安科技深圳有限公司) 2019年 9月 27日 (2019 - 09 - 27) 全文	1-20	X	Y. Wang 等. "style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis" arXiv preprint arXiv:1803.09017, 2018年 3月 23日 (2018 - 03 - 23), 第2-4部分	1-20	A	CN 109616127 A (建湖云飞数据科技有限公司) 2019年 4月 12日 (2019 - 04 - 12) 全文	1-20	A	CN 109754779 A (出门问问信息科技有限公司) 2019年 5月 14日 (2019 - 05 - 14) 全文	1-20	A	US 2018336880 A1 (BAIDU USA LLC等) 2018年 11月 22日 (2018 - 11 - 22) 全文	1-20	A	Ye Jia 等. "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis" arXiv preprint arXiv: 1806.04558v4, 2019年 1月 2日 (2019 - 01 - 02), 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
PX	CN 110288973 A (平安科技深圳有限公司) 2019年 9月 27日 (2019 - 09 - 27) 全文	1-20																					
X	Y. Wang 等. "style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis" arXiv preprint arXiv:1803.09017, 2018年 3月 23日 (2018 - 03 - 23), 第2-4部分	1-20																					
A	CN 109616127 A (建湖云飞数据科技有限公司) 2019年 4月 12日 (2019 - 04 - 12) 全文	1-20																					
A	CN 109754779 A (出门问问信息科技有限公司) 2019年 5月 14日 (2019 - 05 - 14) 全文	1-20																					
A	US 2018336880 A1 (BAIDU USA LLC等) 2018年 11月 22日 (2018 - 11 - 22) 全文	1-20																					
A	Ye Jia 等. "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis" arXiv preprint arXiv: 1806.04558v4, 2019年 1月 2日 (2019 - 01 - 02), 全文	1-20																					
☐ 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 2020年 2月 12日		国际检索报告邮寄日期 2020年 2月 18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																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/117254

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	110288973	A	2019年 9月 27日	无	
CN	109616127	A	2019年 4月 12日	无	
CN	109754779	A	2019年 5月 14日	无	
US	2018336880	A1	2018年 11月 22日	无	