

(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 12 月 17 日 (17.12.2020)



(10) 国际公布号
WO 2020/248393 A1

- (51) 国际专利分类号:
G10L 13/08 (2013.01)
- (21) 国际申请号: PCT/CN2019/103582
- (22) 国际申请日: 2019 年 8 月 30 日 (30.08.2019)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201910515578.3 2019 年 6 月 14 日 (14.06.2019) 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 518033 (CN)。
- (72) 发明人: 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 518033 (CN)。 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 518033 (CN)。

- (74) 代理人: 北京鸿元知识产权代理有限公司 (GRANDER IP LAW FIRM); 中国北京市朝阳区光华路 7 号汉威大厦东区 18A6 室, Beijing 100004 (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,

(54) Title: SPEECH SYNTHESIS METHOD AND SYSTEM, TERMINAL DEVICE, AND READABLE STORAGE MEDIUM

(54) 发明名称: 语音合成方法、系统、终端设备和可读存储介质

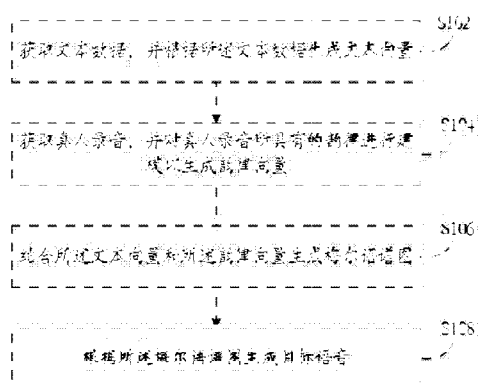


图 1

- S102 Acquire text data, and generate a text vector on the basis of the text data
- S104 Acquire a speech recording of a real person, and model the prosody of the speech recording and generate a prosody vector
- S106 Combine the text vector with the prosody vector and generate a Mel spectrogram
- S108 Generate target speech audio on the basis of the Mel spectrogram

(57) Abstract: A speech synthesis method, comprising: acquiring text data, and generating a text vector on the basis of the text data (S102); acquiring a speech recording of a real person, and modeling the prosody of the speech recording and generating a prosody vector (S104); combining the text vector with the prosody vector and generating a Mel spectrogram (S106); and generating target speech audio on the basis of the Mel spectrogram (S108). In the method, the prosody of the speech recording of a real person is used to implement modeling, and the technique created on the basis of global conditional probability is employed to make the prosody of the synthesized speech more similar to that of the input speech recording of a real person, such that the synthesized speech has high fidelity and high naturalness.

(57) 摘要: 一种语音合成方法, 包括: 获取文本数据, 根据文本数据生成文本向量(S102); 获取真人录音, 对真人录音所具有的韵律进行建模以生成韵律向量(S104); 结合文本向量和韵律向量生成梅尔语谱图(S106); 根据梅尔语谱图生成目标语音(S108)。另外, 本方法还通过真人录音中的韵律进行建模, 并基于全局条件概率生成的方法, 使合成的语音与输入的真人录音具有更为相似的韵律, 进一步使合成语音具有高保真和高自然度的效果。

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))。

语音合成方法、系统、终端设备和可读存储介质

本申请要求于 2019 年 6 月 14 日提交中国专利局，申请号为
201910515578.3、发明名称为“语音合成方法、系统、终端设备和可读存
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

本申请涉及人工智能技术领域，尤其涉及一种语音语义领域，具体涉
及一种语音合成方法、系统、终端设备和可读存储介质。

背景技术

随着科技的发展，机器已经可以通过语音合成技术进行说话。所谓的
语音合成技术，也被称为文语转换技术(Text to Speech, TTS)，其目标
是让机器通过识别和理解，把文本信息变成人造语音输出，是现代人工智
能发展的重要分支。语音合成能够在质量检测、机器问答、残障辅助等领
域发挥极大作用，方便人们的生活。

然而，现有的机器所能合成语音往往具有固定的模式，生成的语音在
韵律方面较为生硬，与真人具有明显的差异，因此在某些对合成语音的拟
人度要求比较高的场景下（如：智能外呼），终端用户往往不能接受如此
生硬的韵律。所以，目前急需一种基于深度学习的语音合成方法。

发明内容

为了解决上述至少一个技术问题，本申请提出了一种语音合成方法、
系统、终端设备和可读存储介质，其能够将真人录音中的韵律转移到合成
的语音中，实现对合成语音保真程度的提升。

为了实现上述目的，本申请第一方面提供了一种语音合成方法，包括：
获取文本数据，并根据所述文本数据生成文本向量；
获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量；
结合所述文本向量和所述韵律向量生成梅尔语谱图；
5 根据所述梅尔语谱图生成目标语音。

本申请第二方面还提供了一种语音合成系统，包括：

文本嵌入模块，用于获取文本数据，并根据所述文本数据生成文本向量；

10 韵律提取模块，用于获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量；

梅尔语谱生成模块，用于结合所述文本向量和所述韵律向量生成梅尔语谱图；

语音生成模块，用于根据所述梅尔语谱图生成目标语音。

15 本申请第三方面还提供了一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如上述语音合成方法的步骤。

本申请第四方面还提供了一种计算机非易失性可读存储介质，所述计算机非易失性可读存储介质中包括计算机程序，所述计算机程序被处理器执行时，实现如上述语音合成方法的步骤。

20 本申请通过获取文本数据和真人录音，并将所述文本数据生成文本向量，再对真人录音所具有的韵律进行建模以生成韵律向量；然后结合所述文本向量和所述韵律向量生成梅尔语谱图；再根据所述梅尔语谱图生成所述目标语音，从而实现将真人录音中的韵律转移到合成的语音中。同时，本申请还通过真人录音中的韵律进行建模，并通过基于全局条件概率生成的方法，使合成的语音与输入的真人录音具有更为相似的韵律，进一步使
25 合成语音具有高保真和高自然度的效果。

本申请的附加方面和优点将在下面的描述部分中给出，部分将从下面的描述中变得明显，或通过本申请的实践了解到。

附图说明

- 5 图 1 示出了本申请一种语音合成方法的流程图。
- 图 2 示出了本申请一个实施例的文本向量生成方法的流程图。
- 图 3 示出了本申请一个实施例的韵律向量生成方法的流程图。
- 图 4 示出了本申请一种语音合成系统的框图。
- 图 5 示出了本申请一个实施例的文本嵌入模块的框图。
- 10 图 6 示出了本申请一个实施例的韵律提取模块的框图。
- 图 7 示出了本申请一种语音合成系统的运行示意图。
- 图 8 示出了本申请一种文本嵌入模块的运行示意图。
- 图 9 示出了本申请一种韵律提取模块的运行示意图。
- 图 10 示出了本申请一种终端设备的示意图。

15

具体实施方式

为了能够更清楚地理解本申请的上述目的、特征和优点，下面结合附图和具体实施方式对本申请进行进一步的详细描述。需要说明的是，在不冲突的情况下，本申请的实施例及实施例中的特征可以相互组合。

- 20 在下面的描述中阐述了很多具体细节以便于充分理解本申请，但是，本申请还可以采用其他不同于在此描述的方式来实施，因此，本申请的保护范围并不受下面公开的具体实施例的限制。

语音合成的主流技术方案有三种：参数合成，波形拼接以及端到端，相较而言，端到端的技术方案能够使生成的语音具有极为卓越的质量。本

25 申请提出的语音合成方法、系统和终端设备，也是基于端到端的技术方案。

图 1 为本申请一种语音合成方法的流程图。

如图 1 所示，本申请的第一方面提供了一种语音合成方法，包括：

S102，获取文本数据，并根据所述文本数据生成文本向量；

S104，获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量；

5 S106，结合所述文本向量和所述韵律向量生成梅尔语谱图；

S108，根据所述梅尔语谱图生成目标语音。

需要说明的是，上述步骤 S106，结合所述文本向量和所述韵律向量生成梅尔语谱图，具体包括：

10 将所述文本向量将作为局部条件，将所述韵律向量作为全局条件，通过所述序列到序列模型映射后，生成所述梅尔语谱图（又称梅尔频谱图）。

进一步的，所述文本向量和所述韵律向量将会被输入一个序列到序列模型（seq2seq）。所述序列到序列模型是一种基于条件概率生成的神经网络模型，输入的文本向量将作为及局部条件，而输入的韵律向量将作为全局条件。最终通过该预训练的序列到序列模型映射后，即可得到梅尔语谱图。
15

需要说明的是，在获取真人录音之后，还包括对所述真人录音进行预增强处理；预增强是以帧为单位进行，目的在于加强高频，增加语音的高频分辨率。因为高频端大约在 800Hz 以上按 6dB/oct（倍频程）衰减，频率越高相应的成分越小，为此要在对真人录音进行分析之前对其高频部分加以提升，也可以改善高频信噪比。
20

图 2 示出了本申请一个实施例的文本向量生成方法的流程图。

如图 2 所示，获取文本数据，并将所述文本数据生成文本向量，具体包括：

S202，获取汉字数据，并对所述汉字数据进行分词处理；

25 S204，将分词处理后的汉字数据转译为带有声调的汉语拼音；

S206，将转译得到的带有声调的汉语拼音转换为一维向量数据；

S208, 按照时间序列将一维向量数据转换为二维向量数据。

需要说明的是, 所述声调包括普通话的第一声、第二声、第三声、第四声和轻声, 用阿拉伯数字 1、2、3、4、5 分别表示普通话的第一声、第二声、第三声、第四声和轻声的声调代码, 但不限于此, 在其它实施例中, 也可以用其他数字表示普通话的四声调和轻声。

图 3 示出了本申请一个实施例的韵律向量生成方法的流程图。

如图 3 所示, 获取真人录音, 并对真人录音所具有的韵律进行建模以生成韵律向量, 具体包括:

S302, 对获取的真人录音进行短时傅里叶变换以得到对应的声谱图;

S304, 对所述声谱图进行梅尔滤波以得到梅尔语谱图;

S306, 对所述梅尔语谱图进行时序上的压缩及特征表示的优化;

S308, 对所述梅尔语谱图进行循环神经网络处理, 并根据时序输出;

S310, 获取每一时刻的输出, 并将所述循环神经网络的全部输出转换为二维的韵律向量。

需要说明的是, 短时傅里叶变换 (Short Time Fourier transform, STFT) 是和傅里叶变换相关的一种数学变换, 用以确定时变信号其局部区域正弦波的频率与相位。具体的, 短时傅里叶变换就是将原来的傅里叶变换在时域截短为多段分别进行傅里叶变换, 每一段记为时刻 t , 对应傅里叶变换求出频域特性, 就可以粗略估计出时刻 t 时的频域特性 (即同时指导了时域和频域的对对应关系)。用于信号截短的工具叫做窗函数 (宽度相当于时间长度), 窗越小, 时域特性越明显, 但是此时由于点数过少导致 FFT 降低了精确度, 导致频域特性不明显。

可以理解, 在其他实施例中, 也可以采用小波变换或 Wigner 分布以得到声谱图, 但不限于此。

具体的, 所述真人录音为一维信号; 所述声谱图为二维信号。

根据本申请的实施例, 在对获取的真人录音进行短时傅里叶变换以得

到对应的声谱图之前，还需要对获取的真人录音添加随机噪声。在进行数据增强，通常会手动合成一些音频，然而某些人工合成(使用软件)的音频有可能会造成一些数字错误，如 underflow、overflow 等。本申请通过在音频中添加随机噪声可以有效避免出现上述数字错误的问题。

5 图 4 示出了本申请一种语音合成系统的框图。

如图 4，本申请第二方面还提供一种语音合成系统 4，所述语音合成系统 4 包括：

文本嵌入模块 41，用于获取文本数据，并根据所述文本数据生成文本向量；

10 韵律提取模块 42，用于获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量；

梅尔语谱生成模块 43，用于结合所述文本向量和所述韵律向量生成梅尔语谱图；

语音生成模块 44，用于根据所述梅尔语谱图生成目标语音。

15 本申请的实施例中，梅尔语谱生成模块 43 为一种序列到序列模型(seq2seq)，所述序列到序列模型是一种基于条件概率生成的神经网络模型。具体的，所述文本向量和所述韵律向量将会被输入一个序列到序列模型，输入的文本向量将作为及局部条件，而输入的韵律向量将作为全局条件。最终通过该预训练的序列到序列模型映射后，即可得到梅尔语谱图。

20 需要说明的是，梅尔语谱生成模块 43 中所使用的序列到序列模型以及所述韵律提取模块使用了同一份不公开的语音数据库进行了联合训练。该语音数据库包含了一位男性/女性说话人(即源说话人)在安静环境下，用专用录音设备录制的总时长约 30 个小时的语音文件，以及每条语音所对应的文本文件。

25 需要说明的是，在获取真人录音之后，还包括对所述真人录音进行预增强处理；预增强是以帧为单位进行，目的在于加强高频，增加语音的高

频分辨率。因为高频端大约在 800Hz 以上按 6dB/oct (倍频程) 衰减, 频率越高相应的成分越小, 为此要在对真人录音进行分析之前对其高频部分加以提升, 也可以改善高频信噪比。

图 5 示出了本申请一个实施例的文本嵌入模块的框图。

如图 5 所示, 所述文本嵌入模块 41 包括:

分词单元 411, 用于获取汉字数据, 并对所述汉字数据进行分词处理;

语言模型单元 412, 用于将分词处理后的汉字数据转译为带有声调的汉语拼音;

独热编码单元 413, 用于将转译得到的带有声调的汉语拼音转换为一维向量数据;

文本向量生成单元 414, 用于按照时间序列将一维向量数据转换为二维向量数据;

其中, 所述声调包括普通话的第一声、第二声、第三声、第四声和轻声, 用阿拉伯数字 1、2、3、4、5 分别表示普通话的第一声、第二声、第三声、第四声和轻声的声调代码。

本申请的实施例中, 独热编码单元执行独热编码 (One-Hot-coding) 的方法为: 使用 N 位状态寄存器来对 N 个状态进行编码, 每个状态都由他独立的寄存器位, 并且在任意时候, 其中只有一位有效。例如对六个状态进行编码:

自然顺序码为 000, 001, 010, 011, 100, 101;

独热编码则是 000001, 000010, 000100, 001000, 010000, 100000。

图 6 示出了本申请一个实施例的韵律提取模块的框图。

如图 6 所示, 所述韵律提取模块 42 包括:

短时傅里叶变换单元 421, 用于对获取的真人录音进行短时傅里叶变换以得到对应的声谱图;

梅尔滤波单元 422, 用于对所述声谱图进行梅尔滤波以得到梅尔语谱

图；

卷积神经网络单元 423，用于对所述梅尔语谱图进行时序上的压缩及特征表示的优化；

GRU 单元 424，用于对所述梅尔语谱图进行循环神经网络处理，并根据
5 时序输出；

韵律向量生成单元 425，用于获取每一时刻的输出，并将所述循环神经网络的全部输出转换为二维的韵律向量。

需要说明的是，短时傅里叶变换（Short Time Fourier transform, STFT）是和傅里叶变换相关的一种数学变换，用以确定时变信号其局部区域正弦波的频率与相位。具体的，短时傅里叶变换就是将原来的傅里叶变换在时域截短为多段分别进行傅里叶变换，每一段记为时刻 t ，对应傅里叶变换求出频域特性，就可以粗略估计出时刻 t 时的频域特性（即同时指导了时域和频域的对对应关系）。用于信号截短的工具叫做窗函数（宽度相当于时间长度），窗越小，时域特性越明显，但是此时由于点数过少导致
10 FFT 降低了精确度，导致频域特性不明显。
15

可以理解，在其他实施例中，短时傅里叶变换单元 421 也可以由小波变换单元或 Wigner 分布单元替代，但不限于此。

具体的，所述真人录音为一维信号；所述声谱图为二维信号。

根据本申请的实施例，在对获取的真人录音进行短时傅里叶变换以得到对应的声谱图之前，还需要对获取的真人录音添加随机噪声。在进行数据增强，通常会手动合成一些音频，然而某些人工合成(使用软件)的音频有可能会造成一些数字错误，如 underflow、overflow 等。本申请通过在音频中添加随机噪声可以有效避免出现上述数字错误的问题。
20

需要说明的是，卷积神经网络(Convolutional Neural Networks, CNN)为一类包含卷积计算且具有深度结构的前馈神经网络，其卷积神经网络包括输入层、隐含层和输出层。
25

卷积神经网络的输入层可以处理多维数据，其中，一维卷积神经网络的输入层接收一维或二维数组，一维数组通常为时间或频谱采样；二维数组可能包含多个通道；二维卷积神经网络的输入层接收二维或三维数组；三维卷积神经网络的输入层接收四维数组。在本申请的实施例中，所述卷积神经网络单元采用二维卷积神经网络，以对真人录音梅尔语谱图进行六层二维卷积操作。

卷积神经网络的隐含层包含卷积层、池化层和全连接层，卷积层的功能是对输入数据进行特征提取，其内部包含多个卷积核，组成卷积核的每个元素都对应一个权重系数和一个偏差量，类似于一个前馈神经网络的神经元；在卷积层进行特征提取后，输出的特征图会被传递至池化层进行特征选择和信息过滤。池化层包含预设定的池化函数，其功能是将特征图中单个点的结果替换为其相邻区域的特征图统计量。全连接层搭建在卷积神经网络隐含层的最后部分，并只向其它全连接层传递信号。

循环神经网络（Recurrent Neural Network, RNN）是一类以序列数据为输入，在序列的演进方向进行递归且所有节点（循环单元）按链式连接形成闭合回路的递归神经网络。循环神经网络具有记忆性、参数共享并且图灵完备，因此能以很高的效率对序列的非线性特征进行学习。

图 7 示出了本申请一种语音合成系统的运行示意图。

如图 7 所示，所述语音合成系统的运行流程为：

向语音合成系统输入需要合成的文本内容（例：您好），文本嵌入模块将所述文本内容嵌入成为文本向量。

向语音合成系统输入真人录音，韵律提取模块将对真人录音所具有的韵律进行建模，形成韵律向量。

将生成的文本向量以及韵律向量输入经过训练的梅尔语谱生成模块生成梅尔语谱图。

使用经过训练的语音生成模块将梅尔语谱图合成为高保真的语音文件，

优选的，所述语音生成模块为语音声码器。

如图 8 所示，在文本嵌入模块中，输入的汉字（您好）首先会进行分词，然后一个经过训练的语言模型会将汉字转译为带有声调的汉语拼音；通过独热编码的方式，将转译得到的拼音字母和声调代码（数字 1~5）转换为一维向量数据，再按照时间序列将其转换为一个二维向量数据。

所述语音生成模块获得了梅尔语谱图后，将梅尔语谱图作为条件输入，生成目标说话人的语音，在本实施例中，所述语音生成模块为 WaveNet 声码器，其由一个非公开的语音数据库训练而成，该数据库与训练梅尔语谱生成模块所用的语音数据库为同一数据库。

如图 9 所示，所述韵律提取模块通过一个循环神经网络实现对真人录音与韵律向量的转换，具体步骤如下：

首先对输入的真人录音进行短时傅里叶变换，再使用梅尔滤波器获得其梅尔语谱图。梅尔语谱图将会被输入一个六层预训练好的卷积神经网络，目的是进行时序上的压缩以及更好的对梅尔语谱中的特征进行表示。处理过的梅尔语谱图会按照时序输入进入基于 GRU 单元的循环神经网络中进行处理，该循环神经网络将根据时序进行输出。在得到每一个时刻的输出后，一个全连接网络会将循环神经网络的输出转化一个二维的向量，此向量就为韵律向量。

图 10 示出了本申请一种终端设备的示意图。

如图 10 所示，本申请第三方面还提供一种终端设备 7，终端设备 7 包括：处理器 71、存储器 72 以及存储在所述存储器 72 中并可在所述处理器 71 上运行的计算机程序 73，例如程序。所述处理器 71 执行所述计算机程序 73 时实现上述各个语音合成方法实施例中的步骤。

本申请的实施例中，所述计算机程序 73 可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器 72 中，并由所述处理器 71 执行，以完成本申请。所述一个或多个模块/单元可以是能够完

成特定功能的一系列计算机程序指令段，该指令段用于描述所述计算机程序 73 在所述终端设备 7 中的执行过程。例如，所述计算机程序 73 可以被分割成文本嵌入模块、韵律提取模块、梅尔语谱生成模块以及语音生成模块，各模块具体功能如下：

- 5 文本嵌入模块，用于获取文本数据，并将所述文本数据生成文本向量；
 韵律提取模块，用于获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量；
 梅尔语谱生成模块，用于结合所述文本向量和所述韵律向量生成梅尔语谱图；

- 10 语音生成模块，用于根据所述梅尔语谱图生成所述目标语音。

所述终端设备 7 可以是桌上型计算机、笔记本、掌上电脑及云端管理服务器等计算设备。所述终端设备 7 可包括，但不仅限于，处理器 71、存储器 72。本领域技术人员可以理解，图 7 仅仅是终端设备 7 的示例，并不构成对终端设备 7 的限定，可以包括比图示更多或更少的部件，或者组合
15 某些部件，或者不同的部件，例如所述终端设备还可以包括输入输出设备、网络接入设备、总线等。

所称处理器 71 可以是中央处理单元(Central Processing Unit, CPU)，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)、
20 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。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。

所述存储器 72 可以是所述终端设备 7 的内部存储单元，例如终端设备
25 7 的硬盘或内存。所述存储器 72 也可以是所述终端设备 7 的外部存储设备，例如所述终端设备 7 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。进一步地, 所述存储器 72 还可以既包括所述终端设备 7 的内部存储单元也包括外部存储设备。所述存储器 72 用于存储所述计算机程序以及所述终端设备所需的其它程序和数据。所述存储器 72 还可以用于暂时地存储已经输出或者将要输出的数据。

本申请第四方面还提供一种计算机非易失性可读存储介质, 所述计算机非易失性可读存储介质中包括计算机程序, 所述计算机程序被处理器执行时, 实现如上述的语音合成方法的步骤。

本申请通过获取文本数据和真人录音, 并将所述文本数据生成文本向量, 再对真人录音所具有的韵律进行建模以生成韵律向量; 然后结合所述文本向量和所述韵律向量生成梅尔语谱图; 再根据所述梅尔语谱图生成所述目标语音, 从而实现将真人录音中的韵律转移到合成的语音中。同时, 本申请还通过真人录音中的韵律进行建模, 并通过基于全局条件概率生成的方法, 使合成的语音与输入的真人录音具有更为相似的韵律, 进一步使合成语音具有高保真和高自然度的效果。

在本申请所提供的几个实施例中, 应该理解到, 所揭露的设备和方法, 可以通过其它的方式实现。以上所描述的设备实施例仅仅是示意性的, 例如, 所述单元的划分, 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, 如: 多个单元或组件可以结合, 或可以集成到另一个系统, 或一些特征可以忽略, 或不执行。另外, 所显示或讨论的各组成部分相互之间的耦合、或直接耦合、或通信连接可以是通过一些接口, 设备或单元的间接耦合或通信连接, 可以是电性的、机械的或其它形式的。

上述作为分离部件说明的单元可以是、或也可以不是物理上分开的, 作为单元显示的部件可以是、或也可以不是物理单元; 既可以位于一个地方, 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全部单元来实现本实施例方案的目的。

另外，在本申请各实施例中的各功能单元可以全部集成在一个处理单元中，也可以是各单元分别单独作为一个单元，也可以两个或两个以上单元集成在一个单元中；上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能单元的形式实现。

5 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：移动存储设备、只读存储器（ROM, Read-Only Memory）、随机存取存储器（RAM, Random Access Memory）、磁碟或者光盘等各种可以存储程序代码的介质。

10 或者，本申请上述集成的单元如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请实施例的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机、服务器、或者网络设备等）执行本申请各个实施例所述方法的全部或部分。而前述的存储介质包括：移动存储设备、ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

15 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，可轻易想到变化或替换，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所述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，其特征在于，包括：

获取文本数据，并根据所述文本数据生成文本向量；

获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量；

结合所述文本向量和所述韵律向量生成梅尔语谱图；

根据所述梅尔语谱图生成目标语音。

2. 根据权利要求 1 所述的语音合成方法，其特征在于，获取文本数据，并将所述文本数据生成文本向量，包括：

获取汉字数据，并对所述汉字数据进行分词处理；

将分词处理后的汉字数据转译为带有声调的汉语拼音；

将转译得到的带有声调的汉语拼音转换为一维向量数据；

按照时间序列将一维向量数据转换为二维向量数据。

3. 根据权利要求 1 所述的语音合成方法，其特征在于，获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量，包括：

对获取的真人录音进行短时傅里叶变换以得到对应的声谱图；

对所述声谱图进行梅尔滤波以得到梅尔语谱图；

对所述梅尔语谱图进行时序上的压缩及特征表示的优化；

对所述梅尔语谱图进行循环神经网络处理，并根据时序输出；

获取每一时刻的输出，并将所述循环神经网络的全部输出转换为二维的韵律向量。

4. 根据权利要求 1 所述的语音合成方法，其特征在于，结合所述文本向量和所述韵律向量生成梅尔语谱图，包括：

将所述文本向量作为局部条件，将所述韵律向量作为全局条件，通过所述序列到序列模型映射后，生成所述梅尔语谱图。

5. 根据权利要求 2 所述的语音合成方法，其特征在于，所述声调包括普通话的第一声、第二声、第三声、第四声和轻声，用阿拉伯数字 1、2、3、4、5 分别表示普通话的第一声、第二声、第三声、第四声和轻声的声调代码。

6. 一种语音合成系统，其特征在于，包括：

文本嵌入模块，用于获取文本数据，并根据所述文本数据生成文本向量；

韵律提取模块，用于获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量；

梅尔语谱生成模块，用于结合所述文本向量和所述韵律向量生成梅尔语谱图；

语音生成模块，用于根据所述梅尔语谱图生成目标语音。

7. 根据权利要求 6 所述的语音合成系统，其特征在于，所述文本嵌入模块包括：

分词单元，用于获取汉字数据，并对所述汉字数据进行分词处理；

语言模型单元，用于将分词处理后的汉字数据转译为带有声调的汉语拼音；

独热编码单元，用于将转译得到的带有声调的汉语拼音转换为一维向量数据；

文本向量生成单元，用于按照时间序列将一维向量数据转换为二维向

量数据。

8. 根据权利要求 7 所述的语音合成系统，其特征在于，所述声调包括普通话的第一声、第二声、第三声、第四声和轻声，用阿拉伯数字 1、2、3、4、5 分别表示普通话的第一声、第二声、第三声、第四声和轻声的声调代码。

9. 根据权利要求 6 所述的语音合成系统，其特征在于，所述韵律提取模块包括：

短时傅里叶变换单元，用于对获取的真人录音进行短时傅里叶变换以得到对应的声谱图；

梅尔滤波单元，用于对所述声谱图进行梅尔滤波以得到梅尔语谱图；

卷积神经网络单元，用于对所述梅尔语谱图进行时序上的压缩及特征表示的优化；

GRU 单元，用于对所述梅尔语谱图进行循环神经网络处理，并根据时序输出；

韵律向量生成单元，用于获取每一时刻的输出，并将所述循环神经网络的全部输出转换为二维的韵律向量。

10. 根据权利要求 6 所述的语音合成系统，其特征在于，所述梅尔语谱生成模块是将所述文本向量作为局部条件，将所述韵律向量作为全局条件，通过所述序列到序列模型映射后，生成所述梅尔语谱图。

11. 一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如下步骤：

获取文本数据，并根据所述文本数据生成文本向量；

获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量；

结合所述文本向量和所述韵律向量生成梅尔语谱图；

根据所述梅尔语谱图生成目标语音。

12. 根据权利要求 11 所述的终端设备，其特征在于，获取文本数据，并将所述文本数据生成文本向量，包括：

获取汉字数据，并对所述汉字数据进行分词处理；

将分词处理后的汉字数据转译为带有声调的汉语拼音；

将转译得到的带有声调的汉语拼音转换为一维向量数据；

按照时间序列将一维向量数据转换为二维向量数据。

13. 根据权利要求 11 所述的终端设备，其特征在于，获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量，包括：

对获取的真人录音进行短时傅里叶变换以得到对应的声谱图；

对所述声谱图进行梅尔滤波以得到梅尔语谱图；

对所述梅尔语谱图进行时序上的压缩及特征表示的优化；

对所述梅尔语谱图进行循环神经网络处理，并根据时序输出；

获取每一时刻的输出，并将所述循环神经网络的全部输出转换为二维的韵律向量。

14. 根据权利要求 11 所述的终端设备，其特征在于，结合所述文本向量和所述韵律向量生成梅尔语谱图，包括：

将所述文本向量作为局部条件，将所述韵律向量作为全局条件，通过所述序列到序列模型映射后，生成所述梅尔语谱图。

15. 根据权利要求 12 所述的终端设备，其特征在于，所述声调包括普通话的第一声、第二声、第三声、第四声和轻声，用阿拉伯数字 1、2、3、4、5 分别表示普通话的第一声、第二声、第三声、第四声和轻声的声调代码。

16. 一种计算机非易失性可读存储介质，其特征在于，所述计算机非易失性可读存储介质中包括计算机程序，所述计算机程序被处理器执行时，实现如下步骤：

获取文本数据，并根据所述文本数据生成文本向量；

获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量；

结合所述文本向量和所述韵律向量生成梅尔语谱图；

根据所述梅尔语谱图生成目标语音。

17. 根据权利要求 16 所述的计算机非易失性可读存储介质，其特征在于，获取文本数据，并将所述文本数据生成文本向量，包括：

获取汉字数据，并对所述汉字数据进行分词处理；

将分词处理后的汉字数据转译为带有声调的汉语拼音；

将转译得到的带有声调的汉语拼音转换为一维向量数据；

按照时间序列将一维向量数据转换为二维向量数据。

18. 根据权利要求 16 所述的计算机非易失性可读存储介质，其特征在于，获取真人录音，并对真人录音所具有的韵律进行建模以生成韵律向量，包括：

对获取的真人录音进行短时傅里叶变换以得到对应的声谱图；

对所述声谱图进行梅尔滤波以得到梅尔语谱图；

对所述梅尔语谱图进行时序上的压缩及特征表示的优化；
对所述梅尔语谱图进行循环神经网络处理，并根据时序输出；
获取每一时刻的输出，并将所述循环神经网络的全部输出转换为二维的韵律向量。

19. 根据权利要求 16 所述的计算机非易失性可读存储介质，其特征在于，结合所述文本向量和所述韵律向量生成梅尔语谱图，包括：

将所述文本向量将作为局部条件，将所述韵律向量作为全局条件，通过所述序列到序列模型映射后，生成所述梅尔语谱图。

20. 根据权利要求 17 所述的计算机非易失性可读存储介质，其特征在于，所述声调包括普通话的第一声、第二声、第三声、第四声和轻声，用阿拉伯数字 1、2、3、4、5 分别表示普通话的第一声、第二声、第三声、第四声和轻声的声调代码。

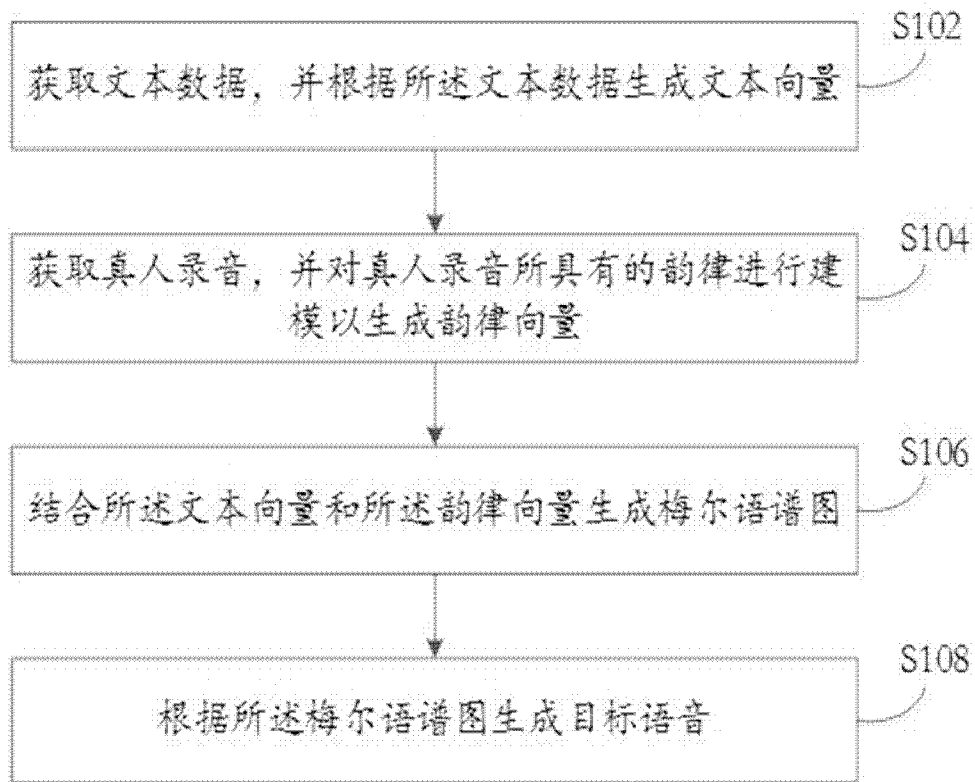


图 1

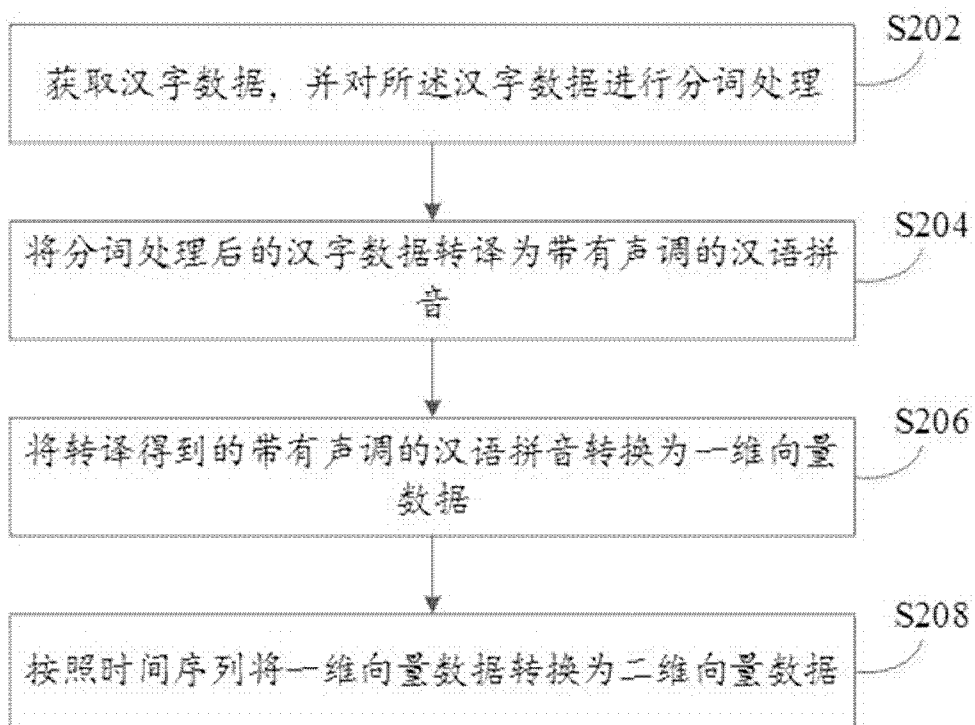


图 2

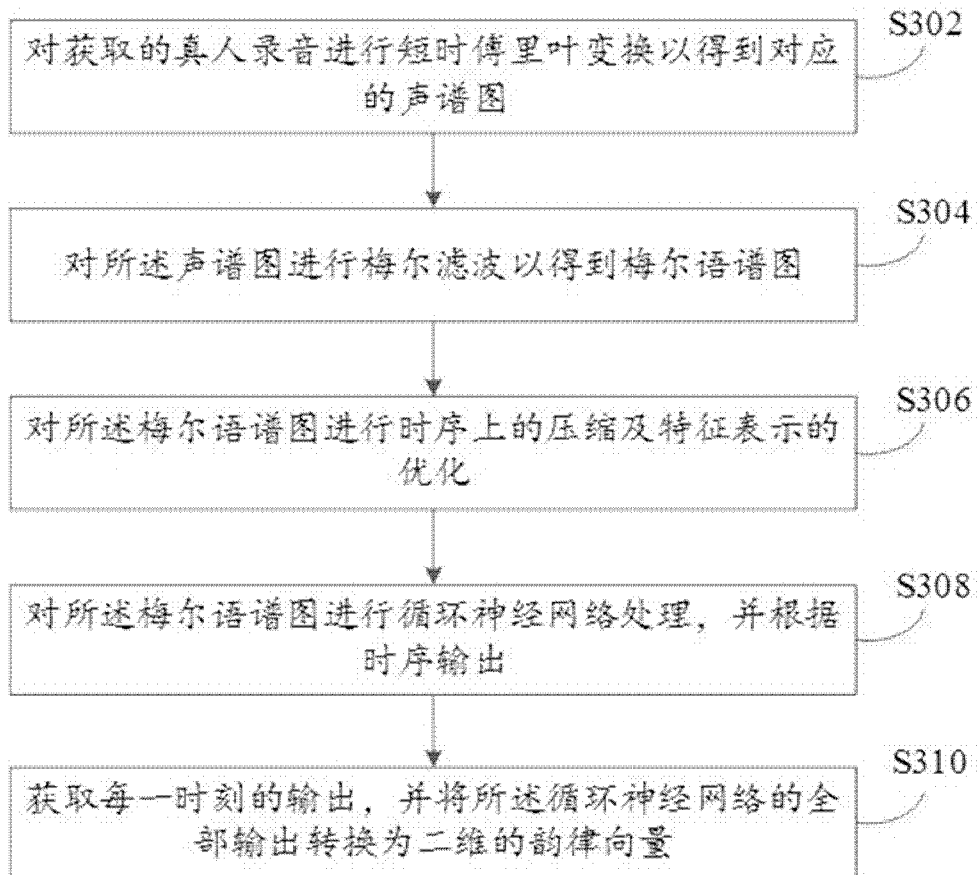


图 3

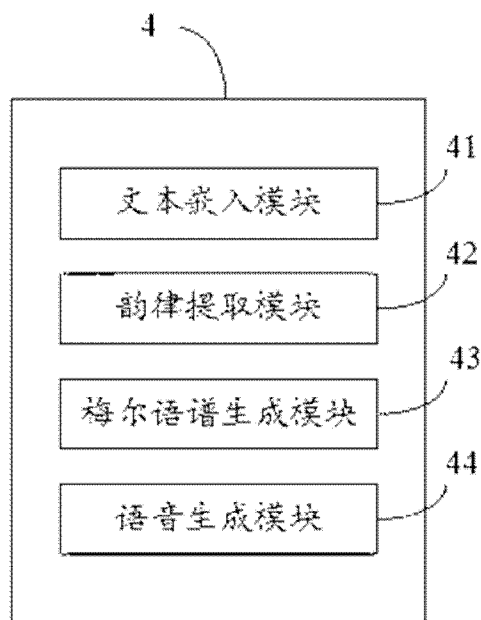


图 4

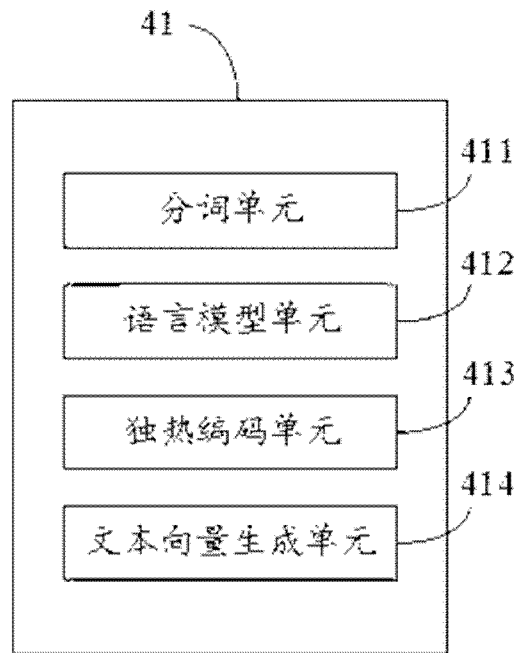


图 5

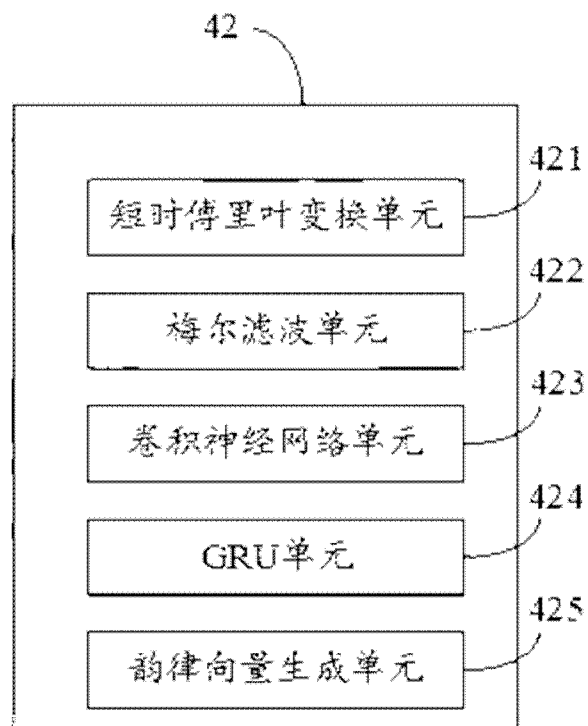


图 6

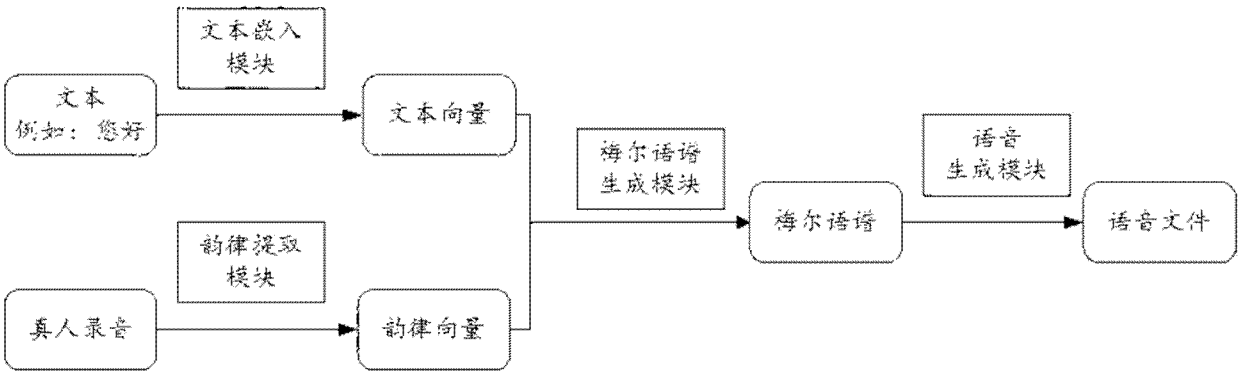


图 7

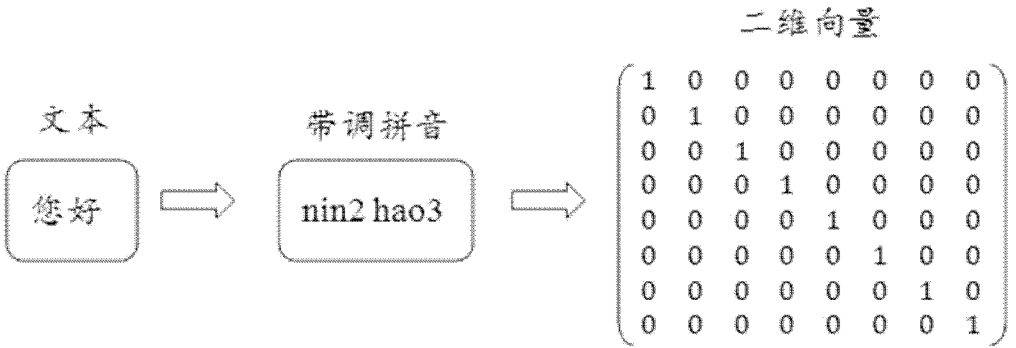


图 8

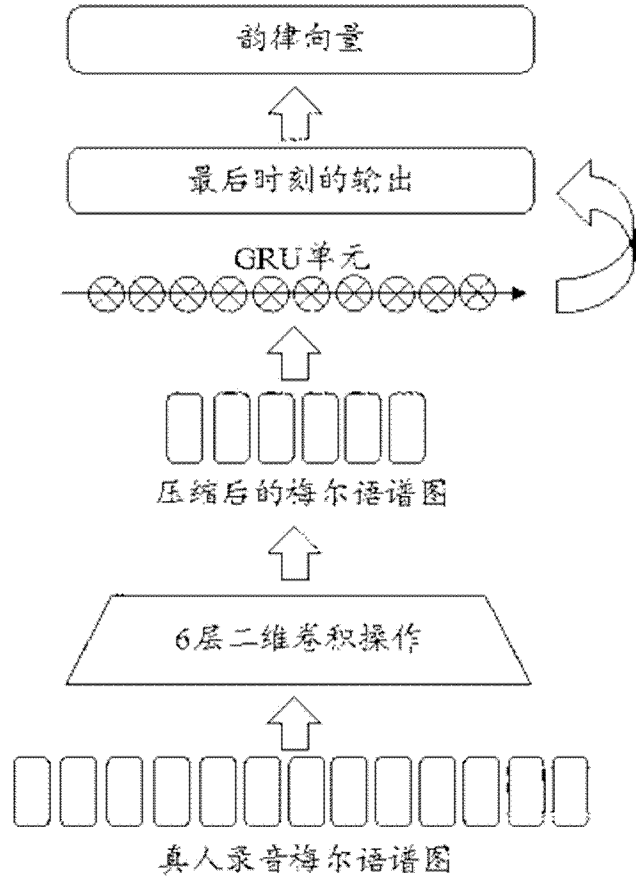


图 9

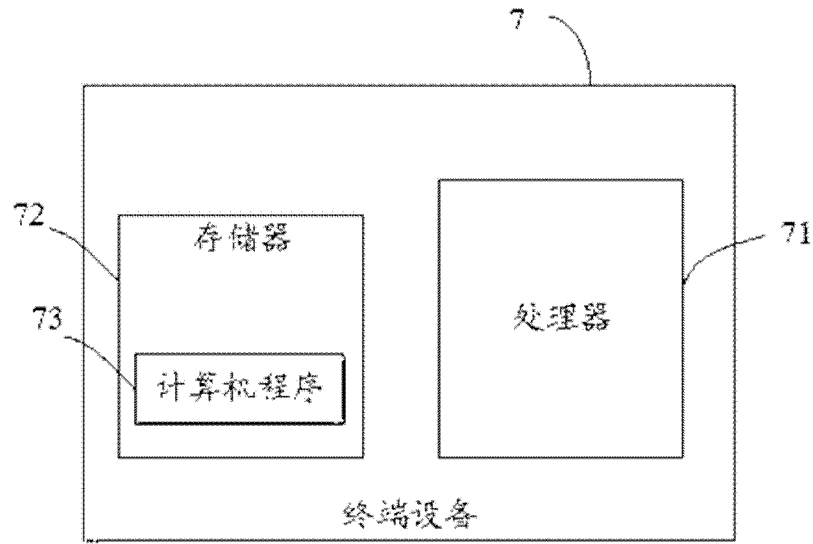


图 10

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/103582

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; CNTXT; VEN; CNKI; 百度学术: 韵律, 端到端, 序列到序列, 语音合成, 语谱, 声谱, 梅尔, prosody, rhythm, end to end, seq2seq, tts, speech synthesis, spectrogram

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	RJ Skerry-Ryan et al. "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron" <i>arXiv preprint arXiv:1803.09047v1</i> , 24 March 2018 (2018-03-24), sections 2-4	1-20
A	CN 109697974 A (BAIDU USA LLC.) 30 April 2019 (2019-04-30) entire document	1-20
A	CN 107452369 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 08 December 2017 (2017-12-08) entire document	1-20
A	CN 109036375 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 18 December 2018 (2018-12-18) entire document	1-20
A	WO 2018183650 A2 (GOOGLE LLC.) 04 October 2018 (2018-10-04) 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

05 March 2020

Date of mailing of the international search report

13 March 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
Information on patent family members

International application No.

PCT/CN2019/103582

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	109697974	A	30 April 2019	US	2019122651	A1	25 April 2019
CN	107452369	A	08 December 2017	US	2019096386	A1	28 March 2019
CN	109036375	A	18 December 2018	WO	2020019885	A1	30 January 2020
WO	2018183650	A2	04 October 2018	EP	3583594	A2	25 December 2019
				AU	2018244917	A1	17 October 2019
				CN	110476206	A	19 November 2019
				KR	20190130634	A	22 November 2019
				WO	2018183650	A3	08 November 2018
				AU	2018244917	B2	05 December 2019
				US	2019311708	A1	10 October 2019
				CA	3058433	A1	04 October 2018

国际检索报告

国际申请号

PCT/CN2019/103582

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; CNTXT; VEN; CNKI; 百度学术: 韵律, 端到端, 序列到序列, 语音合成, 语谱, 声谱, 梅尔, prosody, rhythm, end to end, seq2seq, tts, speech synthesis, spectrogram																				
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>X</td> <td>RJ Skerry-Ryan 等. "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron" arXiv preprint arXiv:1803.09047v1, 2018年 3月 24日 (2018 - 03 - 24), 第2-4部分</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 109697974 A (百度美国有限责任公司) 2019年 4月 30日 (2019 - 04 - 30) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 107452369 A (百度在线网络技术北京有限公司) 2017年 12月 8日 (2017 - 12 - 08) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 109036375 A (腾讯科技深圳有限公司) 2018年 12月 18日 (2018 - 12 - 18) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>WO 2018183650 A2 (GOOGLE LLC.) 2018年 10月 4日 (2018 - 10 - 04) 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	X	RJ Skerry-Ryan 等. "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron" arXiv preprint arXiv:1803.09047v1, 2018年 3月 24日 (2018 - 03 - 24), 第2-4部分	1-20	A	CN 109697974 A (百度美国有限责任公司) 2019年 4月 30日 (2019 - 04 - 30) 全文	1-20	A	CN 107452369 A (百度在线网络技术北京有限公司) 2017年 12月 8日 (2017 - 12 - 08) 全文	1-20	A	CN 109036375 A (腾讯科技深圳有限公司) 2018年 12月 18日 (2018 - 12 - 18) 全文	1-20	A	WO 2018183650 A2 (GOOGLE LLC.) 2018年 10月 4日 (2018 - 10 - 04) 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																		
X	RJ Skerry-Ryan 等. "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron" arXiv preprint arXiv:1803.09047v1, 2018年 3月 24日 (2018 - 03 - 24), 第2-4部分	1-20																		
A	CN 109697974 A (百度美国有限责任公司) 2019年 4月 30日 (2019 - 04 - 30) 全文	1-20																		
A	CN 107452369 A (百度在线网络技术北京有限公司) 2017年 12月 8日 (2017 - 12 - 08) 全文	1-20																		
A	CN 109036375 A (腾讯科技深圳有限公司) 2018年 12月 18日 (2018 - 12 - 18) 全文	1-20																		
A	WO 2018183650 A2 (GOOGLE LLC.) 2018年 10月 4日 (2018 - 10 - 04) 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 3月 5日		国际检索报告邮寄日期 2020年 3月 13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/CN2019/103582

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	109697974	A	2019年 4月 30日	US	2019122651	A1	2019年 4月 25日
CN	107452369	A	2017年 12月 8日	US	2019096386	A1	2019年 3月 28日
CN	109036375	A	2018年 12月 18日	WO	2020019885	A1	2020年 1月 30日
WO	2018183650	A2	2018年 10月 4日	EP	3583594	A2	2019年 12月 25日
				AU	2018244917	A1	2019年 10月 17日
				CN	110476206	A	2019年 11月 19日
				KR	20190130634	A	2019年 11月 22日
				WO	2018183650	A3	2018年 11月 8日
				AU	2018244917	B2	2019年 12月 5日
				US	2019311708	A1	2019年 10月 10日
				CA	3058433	A1	2018年 10月 4日