(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 10 月 8 日 (08.10.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/200178 A1

(51) 国际专利分类号:
G10L 13/02 (2013.01)

(21) 国际申请号: PCT/CN2020/082172

(22) 国际申请日: 2020 年 3 月 30 日 (30.03.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910266289.4 2019年4月3日 (03.04.2019) CN

(71) 申请人: 北京京东尚科信息技术有限公司 (**BEIJING JINGDONG SHANGKE INFORMATION TECHNOLOGY CO.,LTD.**) [CN/CN]; 中国北京市海淀区知春路 76 号 8 层, Beijing 100086 (CN)。 北京京东世纪贸易有限公司 (**BEIJING JINGDONG CENTURY TRADING CO., LTD.**) [CN/CN]; 中国北京市北京经济技术开发区科创十一街18号C座2层201室, Beijing 100176 (CN)。

(72) 发明人: 武执政 (**WU, Zhizheng**); 中国北京市海淀区知春路76号8层, Beijing 100086 (CN)。 张政臣 (**ZHANG, Zhengchen**); 中国北京市海淀区知春路76号8层, Beijing 100086 (CN)。 宋伟 (**SONG, Wei**); 中国北京市海淀区知春路76号8层, Beijing 100086 (CN)。 饶永辉 (**RAO, Yonghui**); 中国北京市海淀区知春路76号8层, Beijing 100086 (CN)。 解知杭 (**XIE, Zhihang**); 中国北京市海淀区知春路76号8层, Beijing 100086 (CN)。 徐光辉 (**XU, Guanghui**); 中国北京市海淀区知春路76号8层, Beijing 100086 (CN)。 刘树勇 (**LIU, Shuyong**); 中国北京市海淀区知春路76号8层, Beijing 100086 (CN)。 马博森 (**MA, Bosen**); 中国北京市海淀区知春路76号8层, Beijing 100086 (CN)。 邱双稳 (**QIU, Shuangwen**); 中国北京市海淀区知春路76号8层, Beijing 100086 (CN)。 林隽民 (**LIN, Junmin**); 中国北京市海淀区知春路76号8层, Beijing 100086 (CN)。

(74) 代理人: 中国国际贸易促进委员会专利商标事务所 (**CCPIT PATENT AND TRADEMARK LAW OFFICE**); 中国北京市西城区阜成门外大街2号万通新世界广场8层, Beijing 100037 (CN)。

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

(54) 发明名称: 语音合成方法、装置和计算机可读存储介质

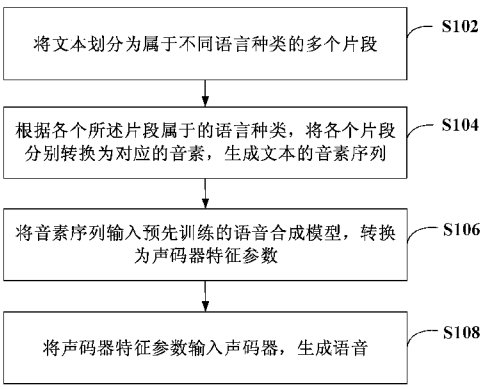

图 1

S102 Divide text into a plurality of segments belonging to different language categories

S104 Convert each segment into a corresponding phoneme according to the language category to which each segment belongs so as to generate a phoneme sequence of the text

S106 Input the phoneme sequence into a pre-trained speech synthesis model and convert same into vocoder feature parameters

S108 Input the vocoder feature parameters into a vocoder to generate speech

(57) **Abstract:** A speech synthesis method, the method comprising: dividing text into a plurality of segments belonging to different language categories (S102); converting each segment into a corresponding phoneme according to the language category to which each segment belongs so as to generate a phoneme sequence of the text (S104); inputting the phoneme sequence into a pre-trained speech synthesis model and converting same into vocoder feature parameters (S106); and inputting the vocoder feature parameters into a vocoder to generate speech (S108).

(57) **摘要**: 一种语音合成方法，该方法包括：将文本划分为属于不同语言种类的多个片段（S102）；根据各个片段属于的语言种类，将各个片段分别转换为对应的音素，生成文本的音素序列（S104）；将音素序列输入预先训练的语音合成模型，转换为声码器特征参数（S106）；将声码器特征参数输入声码器，生成语音（S108）。

WO 2020/200178 A1

[见续页]

(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, 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))。

语音合成方法、装置和计算机可读存储介质

相关申请的交叉引用

本申请是以 CN 申请号为 201910266289.4，申请日为 2019 年 4 月 3 日的申请为基础，并主张其优先权，该 CN 申请的公开内容在此作为整体引入本申请中。

技术领域

本公开涉及计算机技术领域，特别涉及一种语音合成方法、装置和计算机可读存储介质。

背景技术

语音合成系统能够实现文本到语音的转换（Text To Speech，TTS），可以将文本通过一系列的算法操作转换为声音，实现机器模拟人进行发音的过程。

目前的语音合成系统，一般只能支持单独一种语言的发音。

发明内容

发明人发现：目前的语音合成系统一般只支持中文或只支持英文发音，无法实现多种语言的流畅发音。

本公开所要解决的一个技术问题是：如何实现支持多种语言发音的端到端的语音合成系统。

根据本公开的一些实施例，提供的一种语音合成方法，包括：将文本划分为属于不同语言种类的多个片段；根据各个片段属于的语言种类，将各个片段分别转换为对应的音素，生成文本的音素序列；将音素序列输入预先训练的语音合成模型，转换为声码器特征参数；将声码器特征参数输入声码器，生成语音。

在一些实施例中，将文本划分为属于不同语言种类的多个片段包括：根据文本中各个字符的编码，识别各个字符属于的语言种类；将属于同一语言种类的连续字符划分为该语言种类的一个片段。

在一些实施例中，生成文本的音素序列包括：确定文本的韵律结构；根据文本的韵律结构，在与文本中各个字符对应的音素后添加韵律标识，以形成文本的音素序列。

在一些实施例中，将音素序列输入预先训练的语音合成模型，转换为声码器特征

参数包括：将音素序列输入语音合成模型中的声学参数预测模型，转换为声学特征参数；将声学特征参数输入语音合成模型中声码器参数转换模型，得到输出的声码器特征参数。

在一些实施例中，声学参数预测模型包括：编码器、解码器和注意力模型；将音素序列输入语音合成模型中的声学参数预测模型，转换为声学特征参数包括：利用注意力模型，确定当前时刻编码器输出的各个特征表示的注意力权重；判断音素序列中预设元素对应的特征表示的注意力权重是否为各个注意力权重中的最大值，如果是，则结束解码过程。

在一些实施例中，声学特征参数包括语音频谱参数；声码器参数转换模型由多层深度神经网络和长短期记忆网络构成。

在一些实施例中，在声学特征参数的频率小于声码器特征参数的频率的情况下，通过重复声学特征参数进行上采样，使声学特征参数的频率等于声码器特征参数的频率。

在一些实施例中，该方法还包括：训练语音合成模型；其中，训练方法包括：根据预设频率将各个训练文本对应的语音样本划分为不同的帧，并针对每帧提取声学特征参数，分别生成与各个训练文本对应的第一声学特征参数样本；利用各个训练文本和各个训练文本对应的第一声学特征参数样本，对声学参数预测模型进行训练；利用训练完成的声学参数预测模型，将各个训练文本分别转换为第二声学特征参数样本；根据声码器的合成频率，将各个训练文本对应的语音样本分别转换为声码器特征参数样本；利用各个训练文本对应的第二声学特征参数样本和声码器特征参数样本对声码器参数转换模型进行训练。

在一些实施例中，声学参数预测模型包括：编码器、解码器和注意力模型；将音素序列输入语音合成模型中的声学参数预测模型，转换为声学特征参数包括：将音素序列输入编码器，获得编码器输出音素序列中各个元素对应的特征表示；将各个元素对应的特征表示、解码器中第一循环层当前时刻输出的解码器隐状态，以及上一时刻各个元素对应的累积注意力权重信息输入注意力模型，获得上下文向量；将解码器中第一循环层当前时刻输出的解码器隐状态和上下文向量输入解码器的第二循环层，获得解码器第二循环层输出的当前时刻的解码器隐状态；根据解码器输出的各个时刻的解码器隐状态预测声学特征参数。

在一些实施例中，根据各个片段属于的语言种类，将各个片段分别转换为对应的

音素包括：根据各个片段属于的语言种类，将各个片段分别进行文本归一化；根据各个片段属于的语言种类，将归一化后的各个片段分别进行分词；将各个片段的分词，根据各个片段属于的语言种类对应的预设的音素转换表转换为对应的音素；其中，音素包括字符对应的音调。

根据本公开的另一些实施例，提供的一种语音合成装置，包括：语言识别模块，用于将文本划分为属于不同语言种类的多个片段；音素转换模块，用于根据各个片段属于的语言种类，将各个片段分别转换为对应的音素，生成文本的音素序列；参数转换模块，用于将音素序列输入预先训练的语音合成模型，转换为声码器特征参数；语音生成模块，用于将声码器特征参数输入声码器，生成语音。

在一些实施例中，语言识别模块用于根据文本中各个字符的编码，识别各个字符属于的语言种类；将属于同一语言种类的连续字符划分为该语言种类的一个片段。

在一些实施例中，音素转换模块用于确定文本的韵律结构；根据文本的韵律结构，在与文本中各个字符对应的音素后添加韵律标识，以形成文本的音素序列。

在一些实施例中，参数转换模块用于将音素序列输入语音合成模型中的声学参数预测模型，转换为声学特征参数；将声学特征参数输入语音合成模型中声码器参数转换模型，得到输出的声码器特征参数。

在一些实施例中，声学参数预测模型包括：编码器、解码器和注意力模型；参数转换模块用于利用注意力模型，确定当前时刻编码器输出的各个特征表示的注意力权重；判断音素序列中预设元素对应的特征表示的注意力权重是否为各个注意力权重中的最大值，如果是，则结束解码过程。

在一些实施例中，声学特征参数包括语音频谱参数；声码器参数转换模型由多层深度神经网络和长短期记忆网络构成。

在一些实施例中，在声学特征参数的频率小于声码器特征参数的频率的情况下，通过重复声学特征参数进行上采样，使声学特征参数的频率等于声码器特征参数的频率。

在一些实施例中，模型训练模块，用于根据预设频率将各个训练文本对应的语音样本划分为不同的帧，并针对每帧提取声学特征参数，分别生成与各个训练文本对应的第一声学特征参数样本；利用各个训练文本和各个训练文本对应的第一声学特征参数样本，对声学参数预测模型进行训练；利用训练完成的声学参数预测模型，将各个训练文本分别转换为第二声学特征参数样本；根据声码器的合成频率，将各个训练文

本对应的语音样本分别转换为声码器特征参数样本；利用各个训练文本对应的第二声学特征参数样本和声码器特征参数样本对声码器参数转换模型进行训练。

在一些实施例中，声学参数预测模型包括：编码器、解码器和注意力模型；参数转换模块用于将音素序列输入编码器，获得编码器输出音素序列中各个元素对应的特征表示；将各个元素对应的特征表示、解码器中第一循环层当前时刻输出的解码器隐状态，以及上一时刻各个元素对应的累积注意力权重信息输入注意力模型，获得上下文向量；将解码器中第一循环层当前时刻输出的解码器隐状态和上下文向量输入解码器的第二循环层，获得解码器第二循环层输出的当前时刻的解码器隐状态；根据解码器输出的各个时刻的解码器隐状态预测声学特征参数。

在一些实施例中，音素转换模块用于根据各个片段属于的语言种类，将各个片段分别进行文本归一化；根据各个片段属于的语言种类，将归一化后的各个片段分别进行分词；将各个片段的分词，根据各个片段属于的语言种类对应的预设的音素转换表转换为对应的音素；其中，音素包括字符对应的音调。

根据本公开的又一些实施例，提供的一种语音合成装置，包括：存储器；以及耦接至存储器的处理器，处理器被配置为基于存储在存储器中的指令，执行如前述任意实施例的语音合成方法。

根据本公开的再一些实施例，提供的一种计算机可读存储介质，其上存储有计算机程序，其中，该程序被处理器执行时实现前述任意实施例的语音合成方法。

本公开中首先识别文本中的语言种类，将文本划分为属于不同语言种类的多个片段。根据各个片段属于的语言种类，将各个片段分别转换为对应的音素。文本的音素序列被输入语音合成模型转换为声码器特征参数，声码器根据声码器特征参数输出语音。本公开的方案实现了支持多种语言的发音的端到端的语音合成系统。并且根据音素序列转换为声码器特征参数，相对于字符序列直接转换为声码器特征参数，能够使合成的语音更加的准确、流畅和自然。

通过以下参照附图对本公开的示例性实施例的详细描述，本公开的其它特征及其优点将会变得清楚。

附图说明

此处所说明的附图用来提供对本公开的进一步理解，构成本申请的一部分，本公开的示意性实施例及其说明被配置为解释本公开，并不构成对本公开的不当限定。在

附图中：

图 1 示出本公开的一些实施例的语音合成方法的流程示意图。

图 2 示出本公开的一些实施例的语音合成模型的结构示意图。

图 3 示出本公开的另一些实施例的语音合成方法的流程示意图。

图 4 示出本公开的一些实施例的语音合成装置的结构示意图。

图 5 示出本公开的另一些实施例的语音合成装置的结构示意图。

图 6 示出本公开的又一些实施例的语音合成装置的结构示意图。

具体实施方式

下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本公开一部分实施例，而不是全部的实施例。以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本公开及其应用或使用的任何限制。基于本公开中的实施例，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，都属于本公开保护的范围。

本公开提出一种语音合成方法，下面结合图 1 进行描述。

图 1 为本公开语音合成方法一些实施例的流程图。如图 1 所示，该实施例的方法包括：步骤 S102~S108。

在步骤 S102 中，将文本划分为属于不同语言种类的多个片段。

在一些实施例中，根据文本中各个字符的编码，识别各个字符属于的语言种类；将属于同一语言种类的连续字符划分为该语言种类的一个片段。例如，文本中包含中文和英文字符的情况，可以获取文本中字符的 Unicode 码或其他编码，根据 Unicode 码分别识别文本中中文字符和英文字符，进而将文本划分为不同语言的多个片段。如果包含其他语言（例如，日语、法语等）的字符可以根据对应的编码形式进行识别。

下面以文本包含中文和英文为例，描述划分属于不同语言种类的多个片段的具体实施例。（1）根据句子中字符的编码，确定句子中是否存在英文字符，如果不存在执行（2），否则执行（3）。（2）将句子标记为中文句子。（3）确定句子中是否存在中文字符，如果不存在执行（4），否则执行（7）。（4）判断句子是否只包含预设英文字符，预设英文字符可以包括计量单位、缩写和英文编号中至少一项，如果是，执行（5），否则执行（6）。（5）将该句子标记为中文句子。（6）将该句子标记为英文句子。（7）对句子划分中文片段和英文片段。

上述实施例中在句子中只包含预设英文字符的情况下，将句子标记为中文句子，便于后续按照中文将预设的英文字符进行归一化，例如 12km/h 这样的预设英文字符，可以后续进行归一化时转换为 12 千米每小时，后续发出的语音则是中文读法，更加符合中文用户的习惯。本领域技术人员可以理解，参考上述实施例，在句子中只包含一些特殊国际通用字符的情况下，可以根据发音需求将句子标记为预设语言种类，便于后续的文本归一化和语音合成的处理。

上述步骤（7）可以包括以下步骤。（i）判断当前字符的语言种类是否和上一字符的语言种类相同，如果相同，执行（ii），否则执行（iv）。（ii）将当前字符移入当前片段集合。（iii）判断是否到达句尾，如果是，则执行（iv），否则执行（v），（iv）将当前片段集合中的字符标记语言种类，并从当前片段集合移出。（v）将下一字符更新为当前字符，并返回（i）重新开始执行。

在步骤 S104 中，根据各个所述片段属于的语言种类，将各个片段分别转换为对应的音素，生成文本的音素序列。

在一些实施例中，根据各个片段属于的语言种类，将各个片段分别进行文本归一化；根据各个片段属于的语言种类，将归一化后的各个片段分别进行分词；将各个片段的分词，根据该片段属于的语言种类对应的预设的音素转换表转换为对应的音素。文本中通常包含大量的不规范的缩写，例如 12km/s、2019 年等，必须通过归一化操作将这些不规范的文本转换为适合语音合成系统进行语音合成的规范文本。属于不同语言种类的片段需要分别进行文本归一化，可以分别根据不同语言种类的特殊字符对照表，将不规范的字符转换为规范字符，例如，将 12km/s 转换为十二千米每秒，便于后续的音素转换。

由于不同语言的分词方式不同，例如，英文按照单词进行分词，而中文需要根据语义信息等进行分词。因此，根据各个片段属于的语言种类，将各个片段分别进行分词。可以通过查询不同语言种类对应的预设的音素转换表，将各个分词转换为对应的音素（G2P）。一些预设的音素转换表里不存在的单词（OOV），例如拼写错误的单词、新创建的单词、网络单词等，可以通过神经网络等现有技术进行音素转换。预设的音素转换表可以包括多音字的音素对应关系，以便对多音字进行准确的音素转换。也可以通过其他方式识别多音字，或通过其他现有技术进行音素转换，不限于所举示例。

在一些实施例中，音素可以包括字符对应的音调，将音调作为音素的一部分，可

以使合成的语音更加的准确和自然。一些语言例如英语等，没有音调，则不需要在音素序列里添加对应的音调标识。在一些实施例中，还可以对文本划分韵律结构，例如识别文本中的韵律词、韵律短语等。根据文本的韵律结构，在与文本中各个字符对应的音素后添加韵律标识，以形成文本的音素序列。韵律标识可以是韵律词或韵律短语对应的音素后添加的一个表示停顿的特殊标识。韵律结构的预测可以采用现有技术，在此不再赘述。

在步骤 **S106** 中，将音素序列输入预先训练的语音合成模型，转换为声码器特征参数。

根据上述实施例，文本的音素序列可以包括每个字符对应的音素（包括音调）、韵律标识，还可以包括一些特殊符号，例如表示输入的音素序列结束的符号<**EOS**>。语音合成模型的训练过程后续将进行描述。

在一些实施例中，语音合成模型可以包括声学参数预测模型和声码器参数转换模型。声学参数例如包括语音频谱参数，例如，梅尔频谱参数或线性谱参数等。声码器参数根据实际使用的声码器进行确定，例如，声码器采用 **world** 声码器，则声码器参数可以包括基频**(fundamental frequency，F0)**、广义梅尔倒谱系数**(Mel-generalized Cepstral，MGC)**，频带非周期分量**(band a periodical，BAP)**等。将音素序列输入语音合成模型中的声学参数预测模型，可以转换为声学特征参数；将声学特征参数输入语音合成模型中声码器参数转换模型，可以得到输出的声码器特征参数。

声学特征参数预测模型采用 **Encoder-Decoder** 网络结构，包括：编码器、解码器和注意力（**Attention**）模型。输入的音素序列和输出的声学特征参数序列的长度可以是不匹配的，通常声学特征参数序列会比较长。基于 **Encoder-Decoder** 的神经网络结构可以进行灵活的特征预测，符合语音合成的特性。编码器可以包含三层一维卷积和双向 **LSTM**（**Long Short-Term Memory**，长短期记忆网络）。三层一维卷积可以学习得到每个音素的局部上下文信息，双向 **LSTM** 编码则计算得到了每个音素的双向全局信息。编码器模块通过三层一维卷积和双向 **LSTM** 编码能够得到输入音素的非常具有表现力并且包含上下文信息的特征表示。

解码器例如包含两层全连接层和两层 **LSTM**。两层全连接层可以采用 **Dropout** 技术防止神经网络过拟合现象的发生。注意力模型使得解码器在解码过程中可以学习到当前解码时刻需要将注意力关注到哪些输入的音素的内部表示上，通过注意力机制，解码器还可以学习到哪些输入的音素已经完成参数预测，以及当前时刻需要特别关注

哪些音素。注意力模型得到了的编码器的上下文向量，在解码的过程中，通过结合这个上下文向量，可以更好的预测当前时刻需要得到的声学参数以及是否结束解码过程。

在一些实施例中，声学特征参数预测模型中可以执行以下步骤。将音素序列输入编码器，获得编码器输出音素序列中各个元素对应的特征表示。将各个元素对应的特征表示、解码器中第一循环层（例如第一 LSTM）当前时刻输出的解码器隐状态，以及上一时刻各个元素对应的累积注意力权重信息输入注意力模型，获得上下文向量。将解码器中第一循环层当前时刻输出的解码器隐状态和上下文向量输入解码器的第二循环层，获得解码器第二循环层输出的当前时刻的解码器隐状态；根据解码器输出的各个时刻的解码器隐状态预测声学特征参数。例如将解码器隐状态序列进行线性变换得到声学特征参数。

例如，输入音素序列为$\boldsymbol{X}=[\boldsymbol{x_1},\boldsymbol{x_2},\ldots,\boldsymbol{x_j},\ldots\boldsymbol{x_M}]$，编码器输出的特征表示序列为$\boldsymbol{H}=[\boldsymbol{h_1},\boldsymbol{h_2},\ldots,\boldsymbol{h_j},\ldots\boldsymbol{h_M}]$，$\boldsymbol{j}$表示输入音素序列中的各个元素所在的位置，$\boldsymbol{M}$表示音素序列中元素的总个数。解码器输出的隐状态序列为$\boldsymbol{S}=[\boldsymbol{s_1},\boldsymbol{s_2},\ldots,\boldsymbol{s_i},\ldots]$，i 表示解码器输出的时间步骤。音素序列中的韵律标识也会被转换为对应的隐状态，进而转换为解码器隐状态。

例如，上下文向量可以采用以下公式计算。

$$\boldsymbol{e_{i,j}}=\boldsymbol{v^T tanh}(\boldsymbol{Ws_i}+\boldsymbol{Vh_j}+\boldsymbol{Uf_{i,j}}+\boldsymbol{b}) \quad (1)$$

$$\boldsymbol{f_i}=\boldsymbol{F}*\boldsymbol{\alpha_{i-1}} \quad (2)$$

$$\boldsymbol{\beta_i}=\boldsymbol{softmax}(\boldsymbol{e_i}) \quad (3)$$

$$\boldsymbol{c_i}=\sum_{j=0}^{M}\boldsymbol{\beta_{i,j}}*\boldsymbol{h_j} \quad (4)$$

其中，$\boldsymbol{i}$表示的是解码器的时间步骤，$\boldsymbol{j}$表示编码器对应的音素序列中元素的位置，$\boldsymbol{i}$ 和 $\boldsymbol{j}$ 为正整数。$\boldsymbol{v}$，$\boldsymbol{W}$，$\boldsymbol{V}$，$\boldsymbol{U}$，$\boldsymbol{b}$ 是模型训练时学习到的参数，$\boldsymbol{s_i}$表示解码器中第一循环层（例如第一 LSTM）当前第 $\boldsymbol{i}$ 个时刻输出的解码器隐状态。$\boldsymbol{h_j}$表示第 $\boldsymbol{j}$ 个元素对应的特征表示，$\boldsymbol{f_{i,j}}$是$\boldsymbol{f_i}$中的向量，$\boldsymbol{F}$ 是一个预设长度的卷积核，$\boldsymbol{\alpha_{i-1}}$是第 *i-1* 时刻各个元素对应的累积注意力权重信息（Alignments），$\boldsymbol{e_{i,j}}$为数值，$\boldsymbol{e_i}$表示各个元素对应的组成的向量，$\boldsymbol{\beta_i}$为向量，$\boldsymbol{\beta_{i,j}}$表示$\boldsymbol{\beta_i}$中的数值，$\boldsymbol{c_i}$表示第 $\boldsymbol{i}$ 个时刻对应的上下文向量，$\boldsymbol{M}$表示音素序列中元素的总个数。

在一些实施例中，利用所述注意力模型，确定当前时刻编码器输出的各个特征表示的注意力权重；判断音素序列中预设元素对应的特征表示的注意力权重是否为各个注意力权重（即输入音素序列中所有元素对应的注意力权重）中的最大值，如果是，

则结束解码过程。特征表示的注意力权重由注意力模型生成。例如预设元素为音素序列最后一个<EOS>符号。

上述判断是否停止解码的方法，可以使解码器根据实际需求停止解码。通过学习到的 Alignments 信息判断是否需要结束解码过程。如果解码的时候注意力模型已经将注意力转移到了最后符号，但是没有正确的预测结束解码过程，系统可以根据这个 Alignments 信息强制结束解码过程。上述辅助解码结束算法，能够很好的解决模型预测解码过程结束失败或者预测结束不正确的问题，避免声学参数预测模型会继续预测若干帧的声学特征出来，最终合成一些无法理解的语音，提高系统语音输出的准确性、流畅性和自然度。

在预测得到输入音素序列的声学特征参数之后，将声学特征参数（例如梅尔谱参数）输入声码器参数转换模型转换为声码器特征参数，然后就可以通过声码器进行语音合成。

声码器参数转换模型可以采用 DNN-LSTM（深度神经网络-长短期记忆网络）的神经网络结构。该网络结构可以包含多层深度神经网络和长短期记忆网络构成。例如，如图 2 所示，该网络结构包含两层 ReLU（激活函数）连接和一层 LSTM。声学特征参数首先被输入 DNN 网络（例如 ReLU），可以学习声学特征的非线性变换，学习神经网络内部特征表示，相当于一个特征学习的过程。DNN 网络输出的特征被输入 LSTM 学习到声学特征参数的历史依赖信息，以便得到更加平滑的特征转换。发明人通过测试发现，当网络结构包含两层 ReLU 连接和一层 LSTM 时声码器参数转换效果更好。

在一些实施例中，在声学特征参数的频率小于声码器特征参数的频率的情况下，通过重复声学特征参数进行上采样，使声学特征参数的频率等于声码器特征参数的频率。例如，声学参数预测模型以 15ms 为一帧进行参数预测，但是声码器通常以 5ms 为一帧进行语音合成，这样就在时间频率上存在一个不匹配的问题，为了解决两个模型频率不一致的问题，需要将声学参数预测模型的输出进行上采样以匹配声码器模型的频率。可以通过重复声学参数预测模型的输出进行上采样，例如，将声学特征参数重复三次，1*80 维的梅尔谱参数，重复三次可以得到 3*80 维的梅尔谱参数。发明人通过测试确定，相对于学习一个上采样神经网络，或差值等方式进行上采样，通过直接重复特征进行上采样就能够达到很好的效果。

在步骤 S108 中，将声码器特征参数输入声码器，生成语音。

上述实施例中的声码器参数转换模型可以与 world 声码器结合，相对于现有技术中 wavenet（网络结构复杂，无法实时在线生成语音），通过简单的网络架构，可以加快计算速度实现实时语音生成，相对于现有技术中 Griffin-lim 模型，减少了叠音，提高了语音合成的效果。

上述实施例的方法中首先识别文本中的语言种类，将文本划分为属于不同语言种类的多个片段。根据各个片段属于的语言种类，将各个片段分别转换为对应的音素。文本的音素序列被输入语音合成模型转换为声码器特征参数，声码器根据声码器特征参数输出语音。上述实施例的方案实现了支持多种语言的发音的端到端的语音合成系统，并且根据音素序列转换为声码器特征参数，相对于字符序列直接转换为声码器特征参数，能够使合成的语音更加的准确、流畅和自然。进一步通过加入韵律结构、音调等生成音素序列，能够进一步提高语音合成效果。通过新的声码器特征参数转换模型，加快计算速度实现实时语音生成，减少了叠音，进一步提高了语音合成的效果。并且上述实施例中还提出了一种解码器结束方法，可以解决模型预测解码过程结束失败或者预测结束不正确的问题，避免声学参数预测模型最终合成一些无法理解的语音，进一步提高系统语音输出的准确性、流畅性和自然度。

在一些实施例中，训练语音合成模型的方法包括：将每个训练文本对应的语音样本根据声码器的合成频率转换为声码器特征参数样本；将每个训练文本输入待训练的语音合成模型，得到输出的声码器特征参数；将输出的声码器特征参数与对应的声码器特征参数样本进行比对，并根据比对结果调整待训练的语音合成模型的参数，直至完成训练。

为了进一步提高声码器参数转换模型的准确性，下面结合图 3 描述本公开的语音合成模型的训练方法的一些实施例。

图 3 为本公开语音合成方法另一些实施例的流程图。如图 3 所示，该实施例的方法包括：步骤 S302~S310。

在步骤 S302 中，根据预设频率将各个训练文本对应的语音样本划分为不同的帧，并针对每帧提取声学特征参数，分别生成与各个训练文本对应的第一声学特征参数样本。

例如，可以将各个语音样本以 15ms 为一帧的频率进行划分，将每帧样本提取声学特征参数，生成第一声学特征参数样本（例如，梅尔谱参数）。

在步骤 S304 中，利用各个训练文本和各个训练文本对应的第一声学特征参数样

本，对声学参数预测模型进行训练。

可以首先针对每个训练文本，将该训练文本划分为属于不同语言种类的片段，根据各个片段属于的语言种类，将各个片段分别转换为对应的音素，生成该训练文本的音素序列。音素序列可以包括音调、韵律标识等。将各个训练文本的音素序列输入声学参数预测模型，得到输出的与各个训练文本对应的声学特征参数。将同一训练文本对应的输出的声学特征参数与第一声学特征参数样本进行比对，根据比对结果对声学参数预测模型中参数进行调整，直至满足第一预设目标，完成声学参数预测模型的训练。

在步骤 **S306** 中，利用训练完成的声学参数预测模型，将各个训练文本分别转换为第二声学特征参数样本。

将各个训练文本输入训练完成的声学参数预测模型，则可以得到与各个训练文本对应的第二声学特征参数样本。

在步骤 **S308** 中，根据声码器的合成频率，将各个训练文本对应的语音样本分别转换为声码器特征参数样本。

例如，可以将语音样本以 **5ms** 为一帧的频率进行划分，将每帧样本转换为声码器特征参数样本（例如，**MGC**、**BAP**、**log F0**）。步骤 **S308** 的执行顺序不受限制，只要在步骤 **S310** 之前即可。

在步骤 **S310** 中，利用各个训练文本对应的第二声学特征参数样本和声码器特征参数样本对声码器参数转换模型进行训练。

例如，将各个第二声学特征参数样本输入声码器参数转换模型，得到输出的声码器特征参数。将输出的声码器特征参数与对应的声码器特征参数样本进行比对，根据比对结果对声码器参数转换模型中参数进行调整，直至满足第二预设目标，完成声码器参数转换模型的训练。

上述实施例的方法采用声学预测模型预测得到的声学特征参数，作为训练数据进行声码器参数转换模型的训练，可以提高声码器参数转换模型的准确度，使合成的语音更加准确、流畅和自然。这是因为，采用直接在语音文件上提取的真实的声学特征参数（例如，梅尔谱参数）训练声码器参数转换模型，那么在实际进行语音合成的时候就会存在模型的输入特征和训练特征不匹配的差异。具体因为在实际语音合成的过程中，输入的特征是声学参数预测模型预测得到的梅尔谱，声学参数预测模型在解码的过程中，随着解码步数的增加，预测得到的声学特征参数的误差会越来越大，但是

声学参数转换模块训练过程却采用的声音文件真实的声学特征参数，训练得到的模型没有学习过预测得到的声学特征参数以及解码过程中存在误差累积的声学特征参数，所以输入特征和训练特征不匹配会导致声码器参数转换模型性能严重下降。

本公开还提供一种语音合成装置，下面结合图 4 进行描述。

图 4 为本公开语音合成装置的一些实施例的结构图。如图 4 所示，该实施例的装置 40 包括：语言识别模块 402，音素转换模块 404，参数转换模块 406，语音生成模块 408。

语言识别模块 402，将文本划分为属于不同语言种类的多个片段。

在一些实施例中，语言识别模块 402 用于根据文本中各个字符的编码，识别各个字符属于的语言种类；将属于同一语言种类的连续字符划分为该语言种类的一个片段。

音素转换模块 404，用于根据各个片段属于的语言种类，将各个片段分别转换为对应的音素，生成文本的音素序列。

在一些实施例中，音素转换模块 404 用于确定文本的韵律结构；根据文本的韵律结构，在与文本中各个字符对应的音素后添加韵律标识，以形成文本的音素序列。

在一些实施例中，音素转换模块 404 用于根据各个片段属于的语言种类，将各个片段分别进行文本归一化；根据各个片段属于的语言种类，将归一化后的各个片段分别进行分词；将各个片段的分词，根据各个片段属于的语言种类对应的预设的音素转换表转换为对应的音素；其中，音素包括字符对应的音调。

参数转换模块 406，用于将音素序列输入预先训练的语音合成模型，转换为声码器特征参数。

在一些实施例中，参数转换模块 406 用于将音素序列输入语音合成模型中的声学参数预测模型，转换为声学特征参数；将声学特征参数输入语音合成模型中声码器参数转换模型，得到输出的声码器特征参数。

在一些实施例中，声学参数预测模型包括：编码器、解码器和注意力模型；参数转换模块 406 用于利用注意力模型，确定当前时刻编码器输出的各个特征表示的注意力权重；判断音素序列中预设元素对应的特征表示的注意力权重是否为各个注意力权重中的最大值，如果是，则结束解码过程。

在一些实施例中，声学特征参数包括语音频谱参数；声码器参数转换模型由多层深度神经网络和长短期记忆网络构成。

在一些实施例中，在声学特征参数的频率小于声码器特征参数的频率的情况下，

通过重复声学特征参数进行上采样，使声学特征参数的频率等于声码器特征参数的频率。

在一些实施例中，参数转换模块 406 用于将音素序列输入编码器，获得编码器输出音素序列中各个元素对应的特征表示；将各个元素对应的特征表示、解码器中第一循环层当前时刻输出的解码器隐状态，以及上一时刻各个元素对应的累积注意力权重信息输入注意力模型，获得上下文向量；将解码器中第一循环层当前时刻输出的解码器隐状态和上下文向量输入解码器的第二循环层，获得解码器第二循环层输出的当前时刻的解码器隐状态；根据解码器输出的各个时刻的解码器隐状态预测声学特征参数。

语音生成模块 408，用于将声码器特征参数输入声码器，生成语音。

在一些实施例中，如图 4 所示，语音合成装置 40 还包括：模型训练模块 410，用于根据预设频率将各个训练文本对应的语音样本划分为不同的帧，并针对每帧提取声学特征参数，分别生成与各个训练文本对应的第一声学特征参数样本；利用各个训练文本和各个训练文本对应的第一声学特征参数样本，对声学参数预测模型进行训练；利用训练完成的声学参数预测模型，将各个训练文本分别转换为第二声学特征参数样本；根据声码器的合成频率，将各个训练文本对应的语音样本分别转换为声码器特征参数样本；利用各个训练文本对应的第二声学特征参数样本和声码器特征参数样本对声码器参数转换模型进行训练。

本公开的实施例中的语音合成装置可各由各种计算设备或计算机系统来实现，下面结合图 5 以及图 6 进行描述。

图 5 为本公开语音合成装置的一些实施例的结构图。如图 5 所示，该实施例的装置 50 包括：存储器 510 以及耦接至该存储器 510 的处理器 520，处理器 520 被配置为基于存储在存储器 510 中的指令，执行本公开中任意一些实施例中的语音合成方法。

其中，存储器 510 例如可以包括系统存储器、固定非易失性存储介质等。系统存储器例如存储有操作系统、应用程序、引导装载程序（Boot Loader）、数据库以及其他程序等。

图 6 为本公开语音合成装置的另一些实施例的结构图。如图 6 所示，该实施例的装置 60 包括：存储器 610 以及处理器 620，分别与存储器 510 以及处理器 520 类似。还可以包括输入输出接口 630、网络接口 640、存储接口 650 等。这些接口 630，640，650 以及存储器 610 和处理器 620 之间例如可以通过总线 660 连接。其中，输入输出接口 630 为显示器、鼠标、键盘、触摸屏等输入输出设备提供连接接口。网络接口 640

为各种联网设备提供连接接口，例如可以连接到数据库服务器或者云端存储服务器等。存储接口 650 为 SD 卡、U 盘等外置存储设备提供连接接口。

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

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

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

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

以上所述仅为本公开的较佳实施例，并不用以限制本公开，凡在本公开的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本公开的保护范围之内。

权 利 要 求

1. 一种语音合成方法，包括：

将文本划分为属于不同语言种类的多个片段；

根据各个所述片段属于的语言种类，将各个所述片段分别转换为对应的音素，生成所述文本的音素序列；

将所述音素序列输入预先训练的语音合成模型，转换为声码器特征参数；

将所述声码器特征参数输入声码器，生成语音。

2. 根据权利要求 1 所述的语音合成方法，其中，

所述将文本划分为属于不同语言种类的多个片段包括：

根据所述文本中各个字符的编码，识别各个所述字符属于的语言种类；

将属于同一语言种类的连续字符划分为该语言种类的一个片段。

3. 根据权利要求 1 所述的语音合成方法，其中，所述生成所述文本的音素序列包括：

确定所述文本的韵律结构；

根据所述文本的韵律结构，在与所述文本中各个字符对应的音素后添加韵律标识，以形成所述文本的音素序列。

4. 根据权利要求 1 所述的语音合成方法，其中，

所述将所述音素序列输入预先训练的语音合成模型，转换为声码器特征参数包括：

将所述音素序列输入所述语音合成模型中的声学参数预测模型，转换为声学特征参数；

将所述声学特征参数输入所述语音合成模型中声码器参数转换模型，得到输出的声码器特征参数。

5. 根据权利要求 4 所述的语音合成方法，其中，

所述声学参数预测模型包括：编码器、解码器和注意力模型；

所述将所述音素序列输入所述语音合成模型中的声学参数预测模型，转换为声学

特征参数包括：

利用所述注意力模型，确定当前时刻所述编码器输出的各个特征表示的注意力权重；

判断所述音素序列中预设元素对应的特征表示的注意力权重是否为各个注意力权重中的最大值，如果是，则结束解码过程。

6. 根据权利要求 **4** 所述的语音合成方法，其中，

所述声学特征参数包括语音频谱参数；

所述声码器参数转换模型由多层深度神经网络和长短期记忆网络构成。

7. 根据权利要求 **4** 所述的语音合成方法，其中，

在所述声学特征参数的频率小于所述声码器特征参数的频率的情况下，通过重复所述声学特征参数进行上采样，使所述声学特征参数的频率等于所述声码器特征参数的频率。

8. 根据权利要求 **1** 所述的语音合成方法，还包括：训练所述语音合成模型；其中，

所述训练方法包括：

根据预设频率将各个训练文本对应的语音样本划分为不同的帧，并针对每帧提取声学特征参数，分别生成与各个所述训练文本对应的第一声学特征参数样本；

利用各个所述训练文本和各个所述训练文本对应的第一声学特征参数样本，对所述声学参数预测模型进行训练；

利用训练完成的声学参数预测模型，将各个所述训练文本分别转换为第二声学特征参数样本；

根据所述声码器的合成频率，将各个所述训练文本对应的语音样本分别转换为声码器特征参数样本；

利用各个所述训练文本对应的所述第二声学特征参数样本和所述声码器特征参数样本对所述声码器参数转换模型进行训练。

9. 根据权利要求 **4** 所述的语音合成方法，其中，

所述声学参数预测模型包括：编码器、解码器和注意力模型；

所述将所述音素序列输入所述语音合成模型中的声学参数预测模型，转换为声学特征参数包括：

将所述音素序列输入所述编码器，获得所述编码器输出所述音素序列中各个元素对应的特征表示；

将所述各个元素对应的特征表示、所述解码器中第一循环层当前时刻输出的解码器隐状态，以及上一时刻所述各个元素对应的累积注意力权重信息输入所述注意力模型，获得上下文向量；

将所述解码器中第一循环层当前时刻输出的解码器隐状态和所述上下文向量输入所述解码器的第二循环层，获得所述解码器第二循环层输出的当前时刻的解码器隐状态；

根据所述解码器输出的各个时刻的解码器隐状态预测所述声学特征参数。

10. 根据权利要求 **1** 所述的语音合成方法，其中，

所述根据各个所述片段属于的语言种类，将各个所述片段分别转换为对应的音素包括：

根据各个所述片段属于的语言种类，将各个所述片段分别进行文本归一化；

根据各个所述片段属于的语言种类，将归一化后的各个所述片段分别进行分词；

将各个所述片段的分词，根据各个所述片段属于的语言种类对应的预设的音素转换表转换为对应的音素；

其中，音素包括字符对应的音调。

11. 一种语音合成装置，包括：

语言识别模块，用于将文本划分为属于不同语言种类的多个片段；

音素转换模块，用于根据各个所述片段属于的语言种类，将各个所述片段分别转换为对应的音素，生成所述文本的音素序列；

参数转换模块，用于将所述音素序列输入预先训练的语音合成模型，转换为声码器特征参数；

语音生成模块，用于将所述声码器特征参数输入声码器，生成语音。

12. 根据权利要求 **11** 所述的语音合成装置，其中，

所述语言识别模块用于根据所述文本中各个字符的编码，识别各个所述字符属于的语言种类；将属于同一语言种类的连续字符划分为该语言种类的一个片段。

13. 根据权利要求 **11** 所述的语音合成装置，其中，

所述音素转换模块用于确定所述文本的韵律结构；根据所述文本的韵律结构，在与所述文本中各个字符对应的音素后添加韵律标识，以形成所述文本的音素序列。

14. 根据权利要求 **11** 所述的语音合成装置，其中，

所述参数转换模块用于将所述音素序列输入所述语音合成模型中的声学参数预测模型，转换为声学特征参数；将所述声学特征参数输入所述语音合成模型中声码器参数转换模型，得到输出的声码器特征参数。

15. 根据权利要求 **14** 所述的语音合成装置，其中，

所述声学参数预测模型包括：编码器、解码器和注意力模型；

所述参数转换模块用于利用所述注意力模型，确定当前时刻所述编码器输出的各个特征表示的注意力权重；判断所述音素序列中预设元素对应的特征表示的注意力权重是否为各个注意力权重中的最大值，如果是，则结束解码过程。

16. 根据权利要求 **14** 所述的语音合成装置，其中，

所述声学特征参数包括语音频谱参数；

所述声码器参数转换模型由多层深度神经网络和长短期记忆网络构成。

17. 根据权利要求 **14** 所述的语音合成装置，其中，

在所述声学特征参数的频率小于所述声码器特征参数的频率的情况下，通过重复所述声学特征参数进行上采样，使所述声学特征参数的频率等于所述声码器特征参数的频率。

18. 根据权利要求 **11** 所述的语音合成装置，还包括：

模型训练模块，用于根据预设频率将各个训练文本对应的语音样本划分为不同的帧，并针对每帧提取声学特征参数，分别生成与各个所述训练文本对应的第一声学特

征参数样本；利用各个所述训练文本和各个所述训练文本对应的第一声学特征参数样本，对所述声学参数预测模型进行训练；利用训练完成的声学参数预测模型，将各个所述训练文本分别转换为第二声学特征参数样本；根据所述声码器的合成频率，将各个所述训练文本对应的语音样本分别转换为声码器特征参数样本；利用各个所述训练文本对应的所述第二声学特征参数样本和所述声码器特征参数样本对所述声码器参数转换模型进行训练。

19. 根据权利要求 **14** 所述的语音合成装置，其中，

所述声学参数预测模型包括：编码器、解码器和注意力模型；

所述参数转换模块用于将所述音素序列输入所述编码器，获得所述编码器输出所述音素序列中各个元素对应的特征表示；将所述各个元素对应的特征表示、所述解码器中第一循环层当前时刻输出的解码器隐状态，以及上一时刻所述各个元素对应的累积注意力权重信息输入所述注意力模型，获得上下文向量；将所述解码器中第一循环层当前时刻输出的解码器隐状态和所述上下文向量输入所述解码器的第二循环层，获得所述解码器第二循环层输出的当前时刻的解码器隐状态；根据所述解码器输出的各个时刻的解码器隐状态预测所述声学特征参数。

20. 根据权利要求 **11** 所述的语音合成装置，其中，

所述音素转换模块用于根据各个所述片段属于的语言种类，将各个所述片段分别进行文本归一化；根据各个所述片段属于的语言种类，将归一化后的各个所述片段分别进行分词；将各个所述片段的分词，根据各个所述片段属于的语言种类对应的预设的音素转换表转换为对应的音素；

其中，音素包括字符对应的音调。

21. 一种语音合成装置，包括：

存储器；以及

耦接至所述存储器的处理器，所述处理器被配置为基于存储在所述存储器中的指令，执行如权利要求 **1-10** 任一项所述的语音合成方法。

22. 一种计算机可读存储介质，其上存储有计算机程序，其中，该程序被处理器

执行时实现权利要求 1-10 任一项所述方法的步骤。

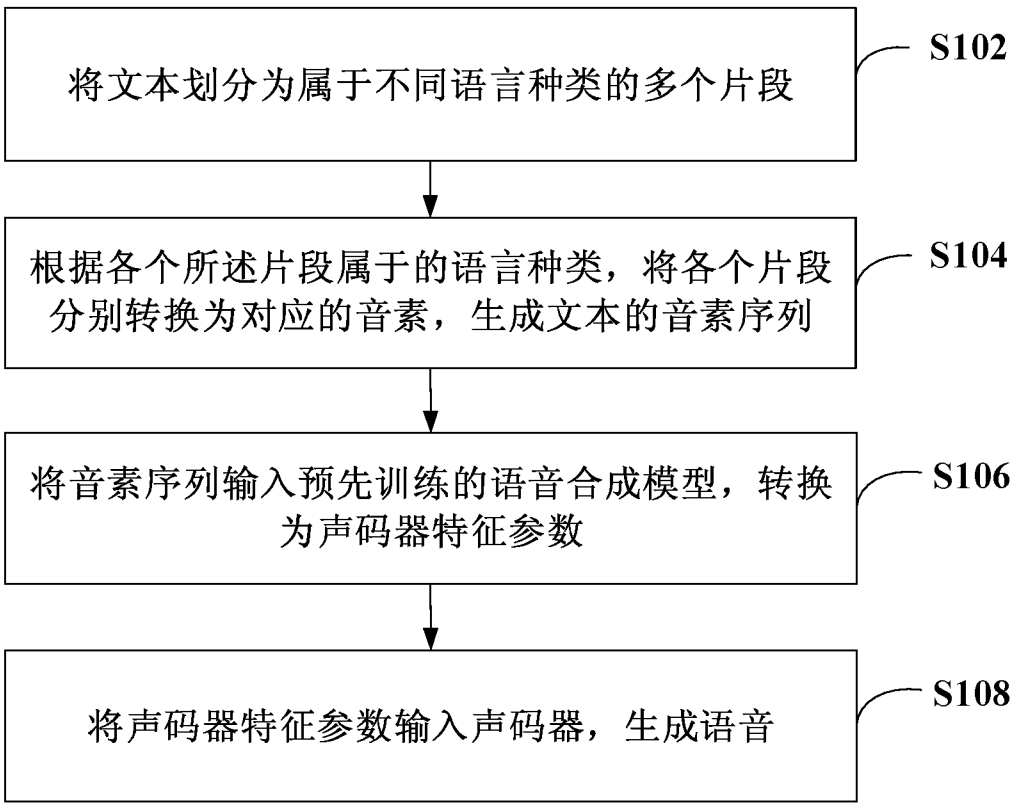
将文本划分为属于不同语言种类的多个片段
S102
根据各个所述片段属于的语言种类，将各个片段分别转换为对应的音素，生成文本的音素序列
S104
将音素序列输入预先训练的语音合成模型，转换为声码器特征参数
S106
将声码器特征参数输入声码器，生成语音
S108

图 1

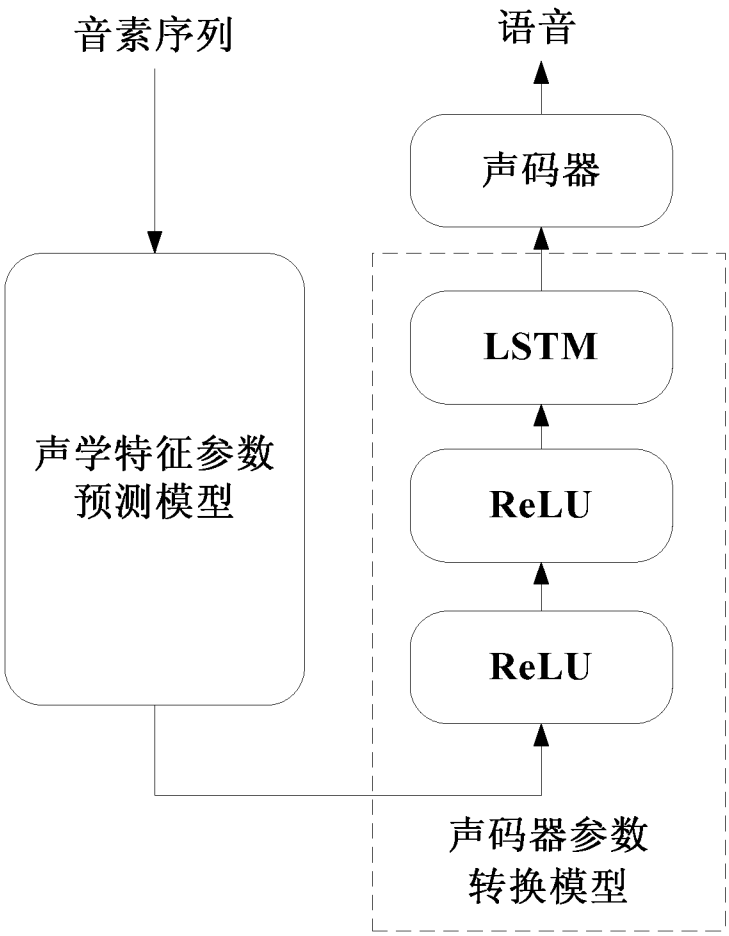

图 2

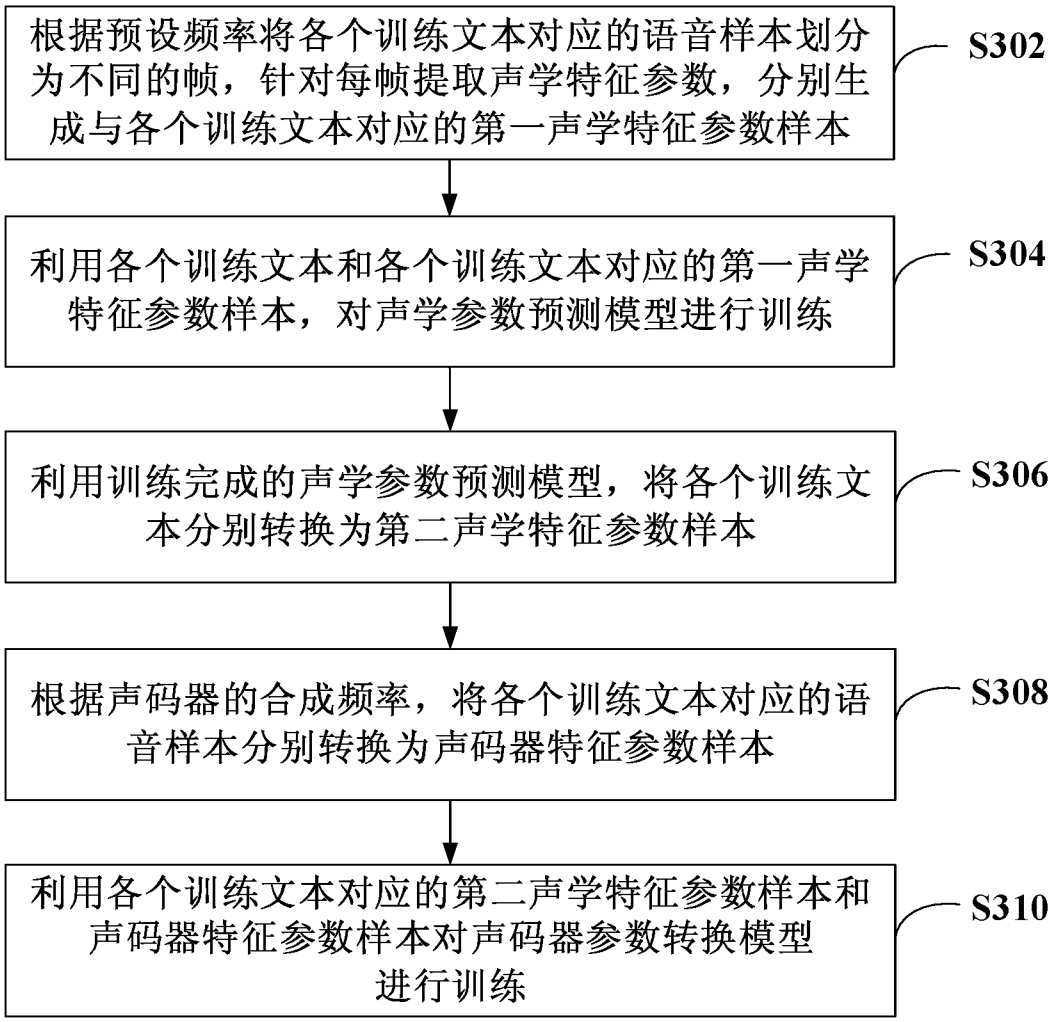
根据预设频率将各个训练文本对应的语音样本划分为不同的帧，针对每帧提取声学特征参数，分别生成与各个训练文本对应的第一声学特征参数样本
S302
利用各个训练文本和各个训练文本对应的第一声学特征参数样本，对声学参数预测模型进行训练
S304
利用训练完成的声学参数预测模型，将各个训练文本分别转换为第二声学特征参数样本
S306
根据声码器的合成频率，将各个训练文本对应的语音样本分别转换为声码器特征参数样本
S308
利用各个训练文本对应的第二声学特征参数样本和声码器特征参数样本对声码器参数转换模型进行训练
S310

图 3


40
语言识别模块
402
音素转换模块
404
参数转换模块
406
语音生成模块
408
模型训练模块
410

图 4

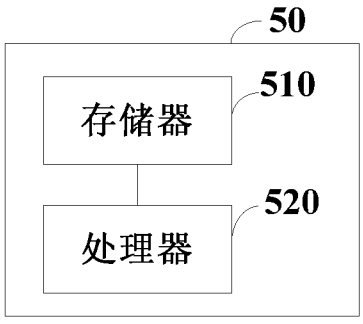
50
510
存储器
520
处理器

图 5

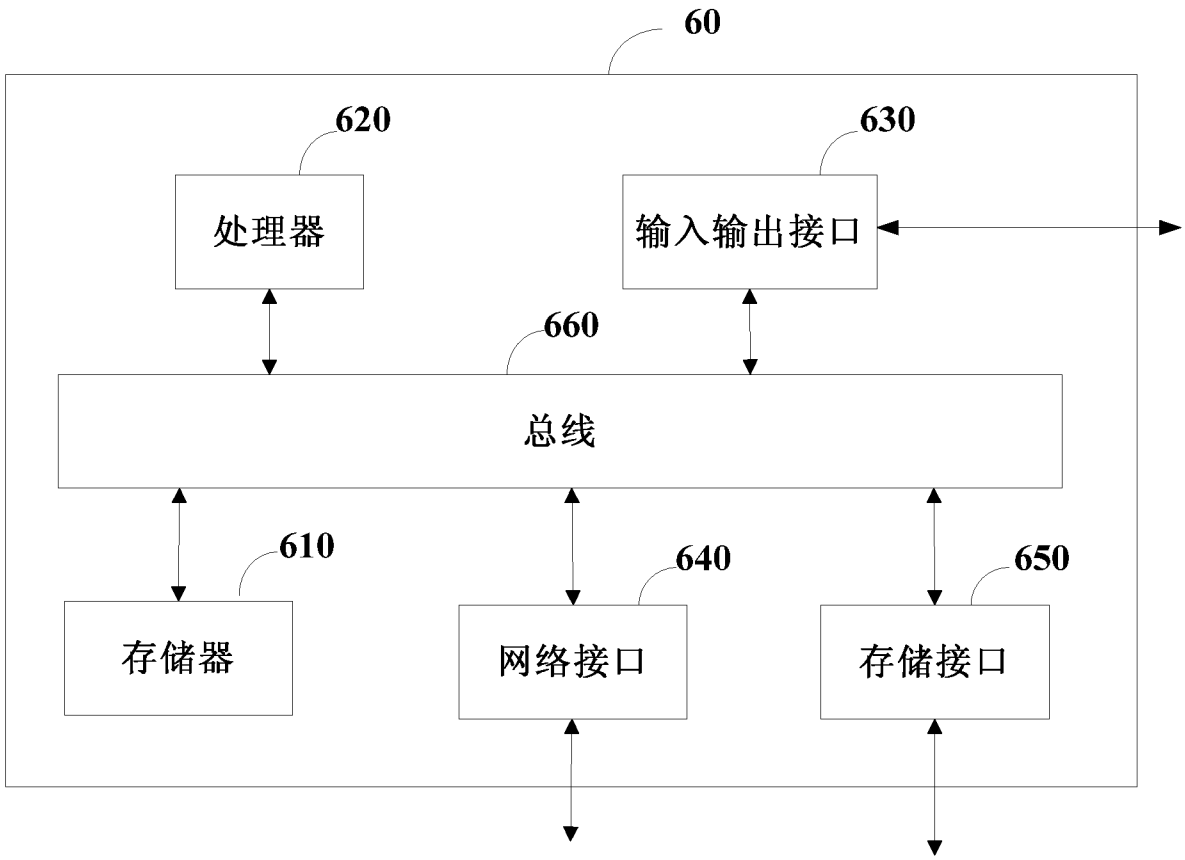

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/082172

A. CLASSIFICATION OF SUBJECT MATTER

G10L 13/02(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)

DWPI; VEN; CNABS; CNTXT; CNKI; 百度学术, IEEE: 多, 混, 语言, 语种, 类型, 种类, 音素, 语音合成, 声码器, 端到端, speech synthesis, text to speech, multilingual, multi, mix, languages, polyglot, vocoder, end to end, seq2seq

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	Liumeng Xue et al. "Building a mixed-lingual neural TTS system with only monolingual data" *arXiv:1904.06063v1,* 12 April 2019 (2019-04-12), sections 2 and 3	1-22
Y	CN 1540625 A (MICROSOFT CORPORATION) 27 October 2004 (2004-10-27) description, page 7, line 12 to page 9, line 17, figures 4-5	1-22
Y	US 2006136216 A1 (DELTA ELECTRONICS INC.) 22 June 2006 (2006-06-22) description, paragraphs [0051]-[0054], figures 2-4	1-22
Y	Eliya Nachmani et al. "UNSUPERVISED POLYGLOT TEXT-TO-SPEECH" *arXiv:1902.02263v1,* 06 February 2019 (2019-02-06), sections 2 and 3	1-22
A	TW 201705019 A (ASUSTEK COMP. INC.) 01 February 2017 (2017-02-01) entire document	1-22
A	US 9484014 B1 (AMAZON TECH. INC.) 01 November 2016 (2016-11-01) entire document	1-22
A	US 2012278081 A1 (CHUN BYUNG HA et al.) 01 November 2012 (2012-11-01) entire document	1-22

☑ 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	Date of mailing of the international search report
29 May 2020	**15 June 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.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106297764 A (IFLYTEK CO., LTD.) 04 January 2017 (2017-01-04) entire document	1-22
A	Bo Li et al. "BYTES ARE ALL YOU NEED:END-TO-END MULTILINGUAL SPEECH RECOGNITION AND SYNTHESIS WITH BYTES" *arXiv:1811.09021v1,* 22 November 2018 (2018-11-22), entire document	1-22

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2020/082172

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	1540625	A	27 October 2004	BR	0400306	A	04 January 2005
				US	2004193398	A1	30 September 2004
				JP	2004287444	A	14 October 2004
				KR	101120710	B1	27 June 2012
				EP	1463031	A1	29 September 2004
				CN	1540625	B	09 June 2010
				US	7496498	B2	24 February 2009
				KR	20040084753	A	06 October 2004
				BR	PI0400306	A	04 January 2005
US	2006136216	A1	22 June 2006	TW	200620240	A	16 June 2006
				TW	I281145	B	11 May 2007
TW	201705019	A	01 February 2017	TW	I605350	B	11 November 2017
				US	9865251	B2	09 January 2018
				US	2017047060	A1	16 February 2017
US	9484014	B1	01 November 2016		None		
US	2012278081	A1	01 November 2012	WO	2010142928	A1	16 December 2010
				GB	2484615	A	18 April 2012
				JP	5398909	B2	29 January 2014
				GB	2484615	B	08 May 2013
				JP	2012529664	A	22 November 2012
				US	8825485	B2	02 September 2014
				GB	201200335	D0	22 February 2012
CN	106297764	A	04 January 2017	CN	106297764	B	30 July 2019

国际检索报告

国际申请号
PCT/CN2020/082172

A. 主题的分类

G10L 13/02(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））

DWPI;VEN;CNABS;CNTXT;CNKI;百度学术，IEEE: 多，混，语言，语种，类型，种类，音素，语音合成，声码器，端到端， speech synthesis,text to speech, multilingual, multi,mix,languages, polyglot, vocoder, end to end, seq2seq

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
PX	Liumeng Xue 等. "Building a mixed-lingual neural TTS system with only mono-lingual data" arXiv:1904.06063v1, 2019年 4月 12日 (2019-04-12), 第2-3部分	1-22
Y	CN 1540625 A (微软公司) 2004年 10月 27日 (2004-10-27) 说明书第7页第12行至第9页第17行，图4-5	1-22
Y	US 2006136216 A1 (DELTA ELECTRONICS INC.) 2006年 6月 22日 (2006-06-22) 说明书第[0051]-[0054]段，附图2-4	1-22
Y	Eliya Nachmani 等. "UNSUPERVISED POLYGLOT TEXT-TO-SPEECH" arXiv:1902.02263v1, 2019年 2月 6日 (2019-02-06), 第2-3部分	1-22
A	TW 201705019 A (ASUSTEK COMP. INC.) 2017年 2月 1日 (2017-02-01) 全文	1-22
A	US 9484014 B1 (AMAZON TECH. INC.) 2016年 11月 1日 (2016-11-01) 全文	1-22

☑ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 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	国际检索报告邮寄日期
2020年 5月 29日	2020年 6月 15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	US 2012278081 A1 (CHUN BYUNG HA等) 2012年 11月 1日 (2012-11-01) 全文	1-22
A	CN 106297764 A (科大讯飞股份有限公司) 2017年 1月 4日 (2017-01-04) 全文	1-22
A	Bo Li 等. "BYTES ARE ALL YOU NEED:END-TO-END MULTILINGUAL SPEECH RECOGNITION AND SYNTHESIS WITH BYTES" arXiv:1811.09021v1, 2018年 11月 22日 (2018-11-22), 全文	1-22

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/082172

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	1540625	A	2004年 10月 27日	BR	0400306	A	2005年 1月 4日
				US	2004193398	A1	2004年 9月 30日
				JP	2004287444	A	2004年 10月 14日
				KR	101120710	B1	2012年 6月 27日
				EP	1463031	A1	2004年 9月 29日
				CN	1540625	B	2010年 6月 9日
				US	7496498	B2	2009年 2月 24日
				KR	20040084753	A	2004年 10月 6日
				BR	PI0400306	A	2005年 1月 4日
US	2006136216	A1	2006年 6月 22日	TW	200620240	A	2006年 6月 16日
				TW	I281145	B	2007年 5月 11日
TW	201705019	A	2017年 2月 1日	TW	I605350	B	2017年 11月 11日
				US	9865251	B2	2018年 1月 9日
				US	2017047060	A1	2017年 2月 16日
US	9484014	B1	2016年 11月 1日		无		
US	2012278081	A1	2012年 11月 1日	WO	2010142928	A1	2010年 12月 16日
				GB	2484615	A	2012年 4月 18日
				JP	5398909	B2	2014年 1月 29日
				GB	2484615	B	2013年 5月 8日
				JP	2012529664	A	2012年 11月 22日
				US	8825485	B2	2014年 9月 2日
				GB	201200335	D0	2012年 2月 22日
CN	106297764	A	2017年 1月 4日	CN	106297764	B	2019年 7月 30日