(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 8 月 27 日 (27.08.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/168752 A1

(51) 国际专利分类号:
G10L 13/02 (2013.01)

(21) 国际申请号: PCT/CN2019/117567

(22) 国际申请日: 2019 年 11 月 12 日 (12.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910135575.7 2019年2月22日 (22.02.2019) 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 518000 (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 518000 (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 518000 (CN)。 肖京 (XIAO, Jing); 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 518000 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司 (SCIHEAD IP LAW FIRM); 中国广东省广州市越秀区先烈中路80号汇华商贸大厦1508室, Guangdong 510070 (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,

(54) **Title:** SPEECH RECOGNITION AND SPEECH SYNTHESIS METHOD AND APPARATUS BASED ON DUAL LEARNING

(54) 发明名称: 一种基于对偶学习的语音识别与语音合成方法及装置

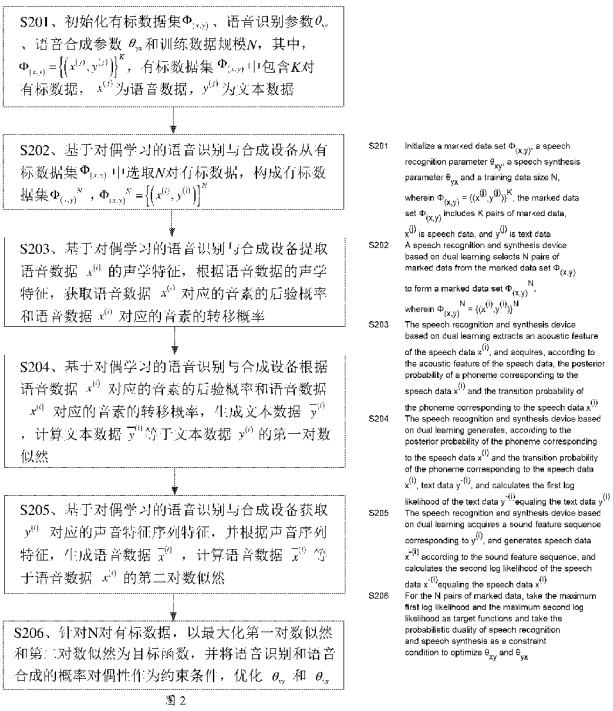

图 2

(57) **Abstract:** A speech recognition and speech synthesis method and apparatus based on dual learning. The method comprises: initializing a marked data set $\Phi_{(x,y)}$, a speech recognition parameter θ_{xy} and a speech synthesis parameter θ_{yx}, wherein $\Phi_{(x,y)}=\{(x^{(j)},y^{(j)})\}^K$, $x^{(j)}$ is speech data, and $y^{(j)}$ is text data; selecting, from $\Phi_{(x,y)}$, N pairs of marked data $\{(x^{(i)},y^{(i)})\}^N$; extracting an acoustic feature of $x^{(i)}$, and according to the acoustic feature of $x^{(i)}$, acquiring the posterior probability of a phoneme corresponding to $x^{(i)}$ and the transition probability

WO 2020/168752 A1

[见续页]

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))。

---

of the phoneme corresponding to $x^{(i)}$ to generate the text data (a), and calculating the first log likelihood of (a) equaling $y^{(i)}$; acquiring a sound feature sequence corresponding to $y^{(i)}$ to generate the speech data (b), and calculating the second log likelihood of (b) equaling $x^{(i)}$; and taking the maximum first log likelihood and the maximum second log likelihood as target functions, and taking the probabilistic duality of speech recognition and speech synthesis as a constraint condition to optimize θ_{xy} and θ_{yx}. According to the method, dual learning is effectively used to perform speech recognition and speech synthesis, thereby increasing the training speed of speech recognition and speech generation and improving the accuracy of an output result.

(57) 摘要：一种基于对偶学习的语音识别与语音合成方法及装置，其中，所述方法包括：初始化有标数据集$\Phi_{(x,y)}$、语音识别参数θ_{xy}、语音合成参数θ_{yx}，$\Phi_{(x,y)}=\{(x^{(j)},y^{(j)})\}^{K}$，$x^{(j)}$为语音数据，$y^{(j)}$为文本数据；从$\Phi_{(x,y)}$中选取N对有标数据$\{(x^{(i)}, y^{(i)})\}^{N}$；提取$x^{(i)}$的声学特征，根据$x^{(i)}$的声学特征，获取$x^{(i)}$对应的音素的后验概率和$x^{(i)}$对应的音素的转移概率生成文本数据（a），计算（a）等于$y^{(i)}$的第一对数似然；获取$y^{(i)}$对应的声音特征序列，生成语音数据（b），计算（b）等于$x^{(i)}$的第二对数似然；以最大化第一对数似然和第二对数似然为目标函数，并将语音识别和语音合成的概率对偶性作为约束条件，优化θ_{xy}和θ_{yx}。该方法有效利用了对偶学习进行语音识别与语音合成，提高语音识别和语音生成的训练速度，以及输出结果的精度。

一种基于对偶学习的语音识别与语音合成方法及装置

本申请要求于 2019 年 02 月 22 日提交中国专利局、申请号为 201910135575.7、申请名称为“ 一种基于对偶学习的语音识别与语音合成方法及装置”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

本申请涉及新能源汽车技术领域，尤其涉及一种基于对偶学习的语音识别与语音合成方法及装置。

背景技术

近年来，以深度学习和增强学习为代表的人工智能技术取得了长足的进步，在很多应用取得了巨大的成功。然而，深度学习限制于大规模的带标注的数据，增强学习限制于持续的可交互环境。首先，大规模的带标注的数据的获取及交互环境的维护的代价均很高，为了使深度学习和增强学习能够取得更广泛的成功，我们需要降低其对大规模标注数据和交互环境的依赖性。为了解决这个问题，出现了一种新的学习范式，我们把它称作对偶学习。

在有监督的学习任务中，发现很多问题具有双重的形式，即输入和输出是以对偶的形式出现的，其中一个任务的输入和输出就是另一个任务的输出和输入，比如在机器翻译里面，不同语言之间的相互翻译互为对偶任务。这两个任务在内部有概率的关系，有相关性模型，但是这种联系没有得到有效的利用，因为人们通常在训练模型时是将两个任务独立地完成的。所以对偶学习的出现就是利用这两个模型之间的相关性，对这两者同时开展训练，简化训练流程，对偶学习并不依赖于大规模标注数据。

传统技术中通常将语音识别和语音生成分开进行训练，未能有效利用语音识别和语音生成之间的对偶性。利用语音识别和语音生成之间的对偶性，联合语音识别训练和语音生成训练进行对偶学习，是语音识别和语音生成技术的一大发展趋势。然而，如何将对偶学习应用于实际场景中依然面临巨大的挑战，如何基于对偶学习有效的进行语音识别和语音生成，提高语音识别和语音生成的训练速度及输出结果的精度是目前急需解决的技术问题。

发明内容

本申请实施例提供了一种基于对偶学习的语音识别与语音合成方法及装置，能够有效的利用对偶学习进行语音识别与语音合成，提高语音识别和语音生成的训练速度，以及提高语音识别和语音生成输出结果的精度。

本申请实施例提供了一种基于对偶学习的语音识别与语音合成方法，该方法包括以下步骤：

初始化有标数据集 $\Phi_{(x,y)}$、语音识别参数 θ_{xy}、语音合成参数 θ_{yx} 和训练数据规模 N，其中，有标数据集 $\Phi_{(x,y)}=\left\{\left(x^{(j)},y^{(j)}\right)\right\}^{K}$，有标数据集 $\Phi_{(x,y)}$ 中包含 K 对有标数据，$\left(x^{(j)},y^{(j)}\right)$ 表示有标数据集 $\Phi_{(x,y)}$ 中的第 j 对有标数据，$x^{(j)}$ 为第 j 对有标数据中的语音数据，$y^{(j)}$ 为第 j 对有标数据中的文本数据，K 为正整数，N 为小于等于 K 的正整数；

从有标数据集 $\Phi_{(x,y)}$ 中选取 N 对有标数据 $\left\{\left(x^{(i)},y^{(i)}\right)\right\}^{N}$；

提取语音数据 $x^{(i)}$ 的声学特征，根据语音数据 $x^{(i)}$ 的声学特征，获取语音数据 $x^{(i)}$ 对应的

音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率；

根据语音数据 $x^{(i)}$ 对应的音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率，生成文本数据 $\bar{y}^{(i)}$，计算文本数据 $\bar{y}^{(i)}$ 等于文本数据 $y^{(i)}$ 的第一对数似然；

获取 $y^{(i)}$ 对应的声音特征序列，并根据声音序列特征，生成语音数据 $\bar{x}^{(i)}$，计算语音数据 $\bar{x}^{(i)}$ 等于语音数据 $x^{(i)}$ 的第二对数似然；

针对 N 对有标数据，以最大化第一对数似然和所述第二对数似然为目标函数，并将语音识别和语音合成的概率对偶性作为约束条件，优化 θ_{xy} 和 θ_{yx}。

本申请实施例还提供了一种基于对偶学习的语音识别与语音合成的装置，该装置能实现上述基于对偶学习的语音识别与语音合成方法所具备的有益效果。其中，该装置的功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。所述硬件或软件包括至少一个与上述功能相对应的模块。可选的，该装置包括初始化单元、选取单元、处理单元、第一生成单元、第二生成单元和优化单元。

初始化单元，用于初始化有标数据集 $\Phi_{(x,y)}$、语音识别参数 θ_{xy}、语音合成参数 θ_{yx} 和训练数据规模 N，其中，有标数据集 $\Phi_{(x,y)}=\left\{\left(x^{(j)},y^{(j)}\right)\right\}^{K}$，有标数据集 $\Phi_{(x,y)}$ 中包含 K 对有标数据，$\left(x^{(j)},y^{(j)}\right)$ 表示有标数据集 $\Phi_{(x,y)}$ 中的第 j 对有标数据，$x^{(j)}$ 为第 j 对有标数据中的语音数据，$y^{(j)}$ 为第 j 对有标数据中的文本数据，K 为正整数，N 为小于等于 K 的正整数。

选取单元，用于从有标数据集 $\Phi_{(x,y)}$ 中选取 N 对有标数据 $\left\{\left(x^{(i)},y^{(i)}\right)\right\}^{N}$。

处理单元，用于提取语音数据 $x^{(i)}$ 的声学特征，根据语音数据 $x^{(i)}$ 的声学特征，获取语音数据 $x^{(i)}$ 对应的音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率。

第一生成单元，用于根据语音数据 $x^{(i)}$ 对应的音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率，生成文本数据 $\bar{y}^{(i)}$，计算文本数据 $\bar{y}^{(i)}$ 等于文本数据 $y^{(i)}$ 的第一对数似然。

第二生成单元，用于获取文本数据 $y^{(i)}$ 对应的声音特征序列，并根据声音序列特征，生成语音数据 $\bar{x}^{(i)}$，计算语音数据 $\bar{x}^{(i)}$ 等于语音数据 $x^{(i)}$ 的第二对数似然。

优化单元，用于针对 N 对有标数据，以最大化第一对数似然和第二对数似然为目标函数，并将语音识别和语音合成的概率对偶性作约束条件，优化 θ_{xy} 和 θ_{yx}。

本申请实施例还提供了一种服务器，该服务器能实现上述基于对偶学习的语音识别与

语音合成方法所具备的有益效果。其中，该服务器的功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。所述硬件或软件包括至少一个与上述功能相对应的模块。该服务器包括存储器、处理器和收发器，存储器用于存储支持服务器执行上述方法的计算机程序，所述计算机程序包括程序指令，处理器用于根据程序指令对服务器的动作进行控制管理，收发器用于支持服务器与其它通信设备的通信。

本申请实施例还提供了一种计算机非易失性可读存储介质，非易失性可读存储介质上存储有指令，当其在处理器上运行时，使得处理器执行上述基于对偶学习的语音识别与语音合成方法。

本申请实施例中，通过获取语音数据$x^{(i)}$对应的音素的后验概率和语音数据$x^{(i)}$对应的音素的转移概率，生成文本数据$\bar{y}^{(i)}$，获取$y^{(i)}$对应的声音特征序列，生成语音数据$\bar{x}^{(i)}$，针对 N 对有标数据，以最大化文本数据$\bar{y}^{(i)}$等于文本数据$y^{(i)}$的对数似然和语音数据$\bar{x}^{(i)}$等于语音数据$x^{(i)}$的对数似然为目标，并将语音识别和语音合成的概率对偶性作为约束条件，从而优化语音识别和语音合成效果。有效的利用了对偶学习进行语音识别与语音合成，提高语音识别和语音生成的训练速度，以及提高语音识别和语音生成输出结果的精度。

本申请附加的方面和优点将在下面的描述中部分给出，这些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

本申请上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中：

图 1 是本申请实施例提供的一种服务器的结构示意图；

图 2 是本申请实施例提供的一种基于对偶学习的语音识别与语音合成方法的流程示意图；

图 3 是本申请实施例提供的一种基于对偶学习的语音识别与语音合成装置的结构示意图。

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描述。应当理解，当在本说明书和所附权利要求书中使用时，术语“包括”和“包含”指示所描述特征、整体、步骤、操作、元素和/或组件的存在，但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。此外，术语“第一”、“第二”和“第三”等是用于区别不同的对象，而并非用于描述特定的顺序。

需要说明的是，在本申请实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本申请。在本申请实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本文中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

对偶学习是一种利用一组对偶任务之间的对偶性建立反馈信号，并用这个信号约束训练的学习方案。对偶性广泛存在于人工智能任务之中，例如，机器翻译就是让机器将自然语言从一种语言翻译到另一种语言，中文到英文和英文到中文互为对偶任务。图像识别和图像合成也互为对偶任务，图像识别指的是给定一张图片，判别它的类别和具体信息。图

像生成指的是给定一个类别和具体信息，生成一张对应的图片。同样，语音识别和语音合成也互为对偶任务，语音识别是让机器通过识别和理解过程把语音信号转变为相应的文本或命令的技术，语音合成是将计算机自己产生的、或外部输入的文字信息，通过机械的、电子的方法转变为语音的技术。

语音识别的应用领域非常广泛，常见的应用系统有：语音输入系统，相对于键盘输入方法，更自然、高效；语音控制系统，即用语音来控制设备的运行，相对于手动控制来说更加快捷、方便；智能对话查询系统，根据客户的语音进行操作，为用户提供自然、友好的数据库检索服务。语音合成技术在我们生活中也具有广泛的应用，如电子阅读、车载语音导航、银行医院排号系统、交通播报等等。本申请实施例提供的基于对偶学习的语音识别与语音合成方法可以应用于终端设备、服务器和车载网络设备等具备语音识别和语音合成功能的网络设备，上述终端设备包括智能手机、智能手环、电子阅读设备、笔记本和平板电脑。本申请对此不做具体限定。下面以服务器为例对上述基于对偶学习的语音识别与语音合成方法的应用设备的功能进行详细介绍。

请参见图 1，图 1 为本申请实施例提供的一种服务器 100 的硬件结构示意图，服务器 100 包括：存储器 101、收发器 102 及与所述存储器 101 和收发器 102 耦合的处理器 103。存储器 101 用于存储计算机程序，所述计算机程序包括程序指令，处理器 103 用于执行存储器 101 存储的程序指令，收发器 102 用于在处理器 103 的控制下与其他设备进行通信。当处理器 103 在执行指令时可根据程序指令执行基于对偶学习的语音识别与语音合成方法。

其中，处理器 103 可以是中央处理器(英文：central processing unit，简称：CPU)，通用处理器，数字信号处理器(英文：digital signal processor，简称：DSP)，专用集成电路(英文：application-specific integrated circuit，简称：ASIC)，现场可编程门阵列(英文：field programmable gate array，简称：FPGA)或者其他可编程逻辑器件、晶体管逻辑器件、硬件部件或者其任意组合。其可以实现或执行结合本申请实施例公开内容所描述的各种示例性的逻辑方框，模块和电路。处理器也可以是实现计算功能的组合，例如包含一个或多个微处理器组合，DSP 和微处理器的组合等等。收发器 102 可以是通信接口、收发电路等，其中，通信接口是统称，可以包括一个或多个接口，例如服务器与终端之间的接口。

可选地，服务器 100 还可以包括总线 104。其中，存储器 101、收发器 102 以及处理器 103 可以通过总线 104 相互连接；总线 104 可以是外设部件互连标准(英文：peripheral component interconnect，简称：PCI)总线或扩展工业标准结构(英文：extended industry standard architecture，简称：EISA)总线等。总线 104 可以分为地址总线、数据总线、控制总线等。为便于表示，图 1 中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

除了图 1 所示的存储器 101、收发器 102、处理器 103 以及上述总线 104 之外，实施例中服务器 100 通常根据该服务器的实际功能，还可以包括其他硬件，对此不再赘述。

在上述运行环境下，本申请实施例提供了如图 2 所示的基于对偶学习的语音识别与语音合成方法。请参阅图 2，所述基于对偶学习的语音识别与语音合成方法包括：

S201、初始化有标数据集 $\Phi_{(x,y)}$、语音识别参数 θ_{xy}、语音合成参数 θ_{yx} 和训练数据规模 N，其中，$\Phi_{(x,y)}=\left\{\left(x^{(j)},y^{(j)}\right)\right\}^{K}$，有标数据集 $\Phi_{(x,y)}$ 中包含 K 对有标数据，$x^{(j)}$ 为语音数据，$y^{(j)}$ 为文本数据。

具体的，选取 K 对有标数据，形成有标数据集 $\Phi_{(x,y)}=\left\{\left(x^{(j)},y^{(j)}\right)\right\}^{K}$，有标数据集 $\Phi_{(x,y)}$

中包含 K 对有标数据，$\left(x^{(j)},y^{(j)}\right)$表示有标数据集$\Phi_{(x,y)}$中的第 j 对有标数据，$x^{(j)}$为第 j 对有标数据中的语音数据，$y^{(j)}$为第 j 对有标数据中的文本数据，K 为正整数，N 为小于等于 K 的正整数。训练数据规模 N 为有标数据集$\Phi_{(x,y)}$中参与基于对偶学习的语音识别与语音合成优化训练的有标数据的数量。语音识别参数θ_{xy}为影响语音识别效果的参数，语音合成参数θ_{yx}为影响语音合成效果的参数。

可以理解，有标数据集$\Phi_{(x,y)}$中的 K 个语音数据的内容均不同，K 个语音数据中的长度可以一致，也可以不一致。语音数据可以来自电视机的新闻播报、日常对话、会议录音等，上述 K 个语音数据的来源场景可以相同，也可以不同。本申请对此均不作具体限定。

可选的，随机初始化语音识别参数θ_{xy}和语音合成参数θ_{yx}，例如θ_{xy}和θ_{yx}的初始值均取 1。

S202、基于对偶学习的语音识别与合成设备从有标数据集$\Phi_{(x,y)}$中选取 N 对有标数据，构成有标数据集$\Phi_{(x,y)}{}^{N}$，$\Phi_{(x,y)}{}^{N}=\left\{\left(x^{(i)},y^{(i)}\right)\right\}^{N}$。

可选的，基于对偶学习的语音识别与合成设备从有标数据集$\Phi_{(x,y)}$中随机选取 N 对有标数据之前，上述基于对偶学习的语音识别与语音合成方法还包括：从有标数据集$\Phi_{(x,y)}$中随机选取 S 对有标数据，对待训练的第一语音识别模型进行预训练，得到经过预训练的第二语音识别模型，以及对待训练的第一语音合成模型进行预训练，得到经过预训练的第二语音合成模型，第二语音识别模型包括深深度神经网络(英文全称：Deep Neural Network，英文缩写：DNN)-隐形马尔科夫模型(英文全称：Hidden Markov Model，英文缩写：HMM)，第二语音合成模型包括编码器、解码器和神经声码器，S 为小于等于 K 的正整数。

可选的，上述基于对偶学习的语音识别与合成设备对待训练的第一语音识别模型进行预训练，得到经过预训练的第二语音识别模型，包括：将上述 S 对有标数据$\left\{\left(x^{(r)},y^{(r)}\right)\right\}^{S}$中的语音数据$x^{(r)}$输入待训练的第一语音识别模型；对语音数据$x^{(r)}$进行预处理，获取语音数据$x^{(r)}$对应的频率倒谱系数特征;利用频率倒谱系数特征对第一语音识别模型的DNN-HMM模型进行预训练，得到第二语音识别模型，第二语音识别模型包括训练后的 DNN-HMM 模型。

可选的，上述基于对偶学习的语音识别与合成设备对待训练的第一语音识别模型进行预训练，得到经过预训练的第二语音识别模型，具体包括：基于对偶学习的语音识别与合成设备将有标数据$\left\{\left(x^{(r)},y^{(r)}\right)\right\}^{S}$中的语音数据$x^{(r)}$输入待训练的第一语音识别模型，首先，对上述语音数据$x^{(r)}$进行预处理，获取语音数据$x^{(r)}$对应的频率倒谱系数特征。然后，基于

对偶学习的语音识别与合成设备将上述频率倒谱系数特征作为输入数据，对由高斯混合模型(英文全称：Adaptive background mixture models for real-time tracking，英文缩写：GMM)和隐形马尔科夫模型(英文全称：Hidden Markov Model，英文缩写：HMM)构成声学模型进行训练，获取预训练后的 GMM 输出的音素状态的似然概率特征及预训练后的 HMM 输出的音素状态的转移概率。基于对偶学习的语音识别与合成设备通过强制对齐将音素状态的似然概率特征转换为音素状态的后验概率特征，根据上述 S 对有标数据及音素状态的后验概率特征，得到深度神经网络(英文全称：Deep Neural Network，英文缩写：DNN)模型中输出层节点间的矩阵权重值和矩阵偏置值，生成预训练后的 DNN 模型。第二语音识别模型包括上述预训练后的 DNN 模型和上述预训练后的 HMM。

可选的，上述基于对偶学习的语音识别与合成设备对待训练的第一语音合成模型进行预训练，得到经过预训练的第二语音合成模型，包括：将上述 S 对有标数据 $\left\{\left(x^{(r)}, y^{(r)}\right)\right\}^{S}$ 中的文本数据 $y^{(r)}$ 输入待训练的第一语音合成模型；利用文本数据 $y^{(r)}$ 对第一语音合成模型的编码器、解码器、神经声码器进行预训练，得到第二语言合成模型，第二语言合成模型包括训练后的编码器、训练后的解码器和训练后的神经声码器。

可选的，上述基于对偶学习的语音识别与合成设备对待训练的第一语音识别模型进行预训练，得到经过预训练的第二语音识别模型，具体包括：基于对偶学习的语音识别与合成设备从有标数据集 $\Phi_{(x,y)}$ 中随机选取 S 对有标数据 $\left\{\left(x^{(t)}, y^{(t)}\right)\right\}^{S}$ 对待训练的第一语音合成模型进行预训练，得到经过预训练的第二语音合成模型。具体包括以下步骤：基于对偶学习的语音识别与合成设备将有标数据 $\left\{\left(x^{(t)}, y^{(t)}\right)\right\}^{S}$ 中的文本数据 $y^{(t)}$ 输入待训练的第一语音合成模型，首先，通过编码器对文本数据进行文本解析，获取文本数据 $y^{(t)}$ 对应的表示文本语义的中间语义向量。然后，基于对偶学习的语音识别与合成设备将上述中间语义向量输入解码器，获取文本数据 $y^{(t)}$ 对应的声音序列特征。将上述声音序列特征输入神经声码器，输出文本数据 $y^{(t)}$ 对应的语音数据。上述编码器、解码器和神经声码器均采用循环神经网络(英文全称：Hidden Markov Model，英文缩写：HMM)模型，第二语音合成模型包括上述编码器、解码器和神经声码器。

可以理解，GMM 就是用高斯概率密度函数精确地量化事物，它是一个将事物分解为若干的基于高斯概率密度函数形成的模型。HMM 是关于时序的概率模型，描述由一个隐藏的马尔可夫链随机生成不可观测的状态随机序列，再由各个状态生成一个观测而产生观测随机序列的过程。

S203、基于对偶学习的语音识别与合成设备提取语音数据 $x^{(i)}$ 的声学特征，根据语音数据 $x^{(i)}$ 的声学特征，获取语音数据 $x^{(i)}$ 对应的音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率。

具体的，基于对偶学习的语音识别与合成设备将语音数据 $x^{(i)}$ 输入第二语音识别模型，滤除不重要的信息以及背景噪声，将语音数据 $x^{(i)}$ 分成多帧语音信号。对每帧语音信号进行

分析处理，提取语音数据 $x^{(i)}$ 对应的每帧语音信号的滤波器组特征作为语音数据 $x^{(i)}$ 的声学特征。基于对偶学习的语音识别与合成设备将语音数据 $x^{(i)}$ 的声学特征输入第二语音识别模型中的 DNN 模型，获取 DNN 模型输出的语音数据 $x^{(i)}$ 对应的音素的后验概率，并将语音数据 $x^{(i)}$ 对应的音素输入第二语音识别模型中的 HMM，获取语音数据 $x^{(i)}$ 对应的音素的转移概率。

可以理解，HMM 输出的音素转移概率包括第一音素状态转至第一音素状态的音素转移概率和第一音素状态转至第二音素状态的音素转换概率，第二音素状态为第一音素状态的下一状态。

S204、基于对偶学习的语音识别与合成设备根据语音数据 $x^{(i)}$ 对应的音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率，生成文本数据 $\bar{y}^{(i)}$，计算文本数据 $\bar{y}^{(i)}$ 等于文本数据 $y^{(i)}$ 的第一对数似然。

可选的，基于对偶学习的语音识别与合成设备根据语音数据 $x^{(i)}$ 对应的音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率确定 w 条网络路径的概率；获取上述 w 条网络路径中概率最大的网络路径对应的文本数据 $\bar{y}^{(i)}$，w 为大于零的正整数。

具体的，基于对偶学习的语音识别与合成设备根据语音数据 $x^{(i)}$ 对应的音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率得到不同单词的概率，不同单词组成不同的网络路径，获取每条网络路径的概率，选择概率最大的网络路径作为最优的网络路径，根据上述最优的网络路径生成相应的文本数据 $\bar{y}^{(i)}$。

HMM 模型是对语音信号的时间序列结构建立统计模型，可以看作一个数学上的双重随机过程:一个是用具有有限状态数的马尔可夫链来模拟语音信号统计特性变化的隐含随机过程，另一个是与马尔可夫链链的每一个状态相关联的外界可见的观测序列的随机过程。HMM 模型包含以下元素：隐藏状态、观察序列、隐藏状态的初始概率分布、隐藏状态的转移概率矩阵，观测值的发射概率。语音识别过程中，给定一个训练好的 HMM 模型和一个观测序列（即语音数据的声学特征），找到观测序列对应的最优的状态序列，从而将语音转化为文本。根据每个单词的发音过程，以音素作为隐藏节点，音素的变化过程构成了 HMM 状态序列，每一个音素以一定的概率密度函数生成观测向量。

可以理解，根据每个单词的 HMM 状态转移概率计算每个状态生成该观测值的概率，若一个单词的 HMM 状态序列的联合概率最大，则判断该段语音对应上述单词。举例来说，以单词 five 的语音数据为例，单词“five”由 [f]、[ay]和[v]这三个音素状态连接而成，隐藏节点的每一个状态对应于一个单独音素。我们以单词“one”、“two”、“three”和“five”为例，采用前向算法分别计算观测序列的后验概率，并从中去找概率最大的单词作为识别结果。

可选的，计算$\bar{y}^{(i)}$等于$y^{(i)}$的第一对数似然，即计算将$x^{(i)}$输入第二语音识别模型，识别出$y^{(i)}$的第一对数似然。第一对数似然表征条件概率分布$P_f\left(y^{(i)}|x^{(i)},\theta_{xy}\right)$的对数似然函数，第一对数似然的计算表达式如下所示。

$$logP_f\left(y^{(i)}|x^{(i)},\theta_{xy}\right)=logP\left\{f\left(x^{(i)}\right)=y^{(i)},\theta_{xy}\right\}$$

S205、基于对偶学习的语音识别与合成设备获取文本数据$y^{(i)}$对应的声音特征序列特征，并根据声音序列特征，生成语音数据$\bar{x}^{(i)}$，计算语音数据$\bar{x}^{(i)}$等于语音数据$x^{(i)}$的第二对数似然。

基于对偶学习的语音识别与合成设备获取$y^{(i)}$对应的声音特征序列特征，并根据声音序列特征，生成语音数据$\bar{x}^{(i)}$。具体的，基于对偶学习的语音识别与合成设备将文本数据$y^{(i)}$输入第二语音合成模型，首先，将文本数据$y^{(i)}$拆分成具有语义的最小单位词。将文本数据$y^{(i)}$对应的最小单位词输入第二语音合成模型的编码器，对文本数据$y^{(i)}$对应的最小单位词进行语义分析并分类。然后，基于对偶学习的语音识别与合成设备对文本数据$y^{(i)}$对应的最小单位词进行分类别的编码，输出文本数据$y^{(i)}$对应的具有固定长度的中间语义向量。将上述中间语义向量输入第二语音合成模型的解码器，基于对偶学习的语音识别与合成设备对上述中间语义向量进行语义分析，并生成文本数据$y^{(i)}$对应的声音序列特征。将上述声音序列特征输入第二语音合成模型的神经声码器，生成语音数据$\bar{x}^{(i)}$。

可选的，对文本数据$y^{(i)}$对应的最小单位词进行语义分析并分类，所分类别包括：中文、英文、韩文、数字、拼音和地名等。针对不同类别的最小单位词有不同的编码规则。

可选的，计算$\bar{x}^{(i)}$等于$x^{(i)}$的第二对数似然，即计算将$y^{(i)}$输入第二语音合成模型，识别出$x^{(i)}$的第二对数似然。第二对数似然的计算表达式如下所示。

$$logP_f\left(x^{(i)}|y^{(i)},\theta_{yx}\right)=logP\left\{g\left(y^{(i)}\right)=x^{(i)},\theta_{yx}\right\}$$

S206、基于对偶学习的语音识别与合成设备针对N对有标数据，以最大化第一对数似然和第二对数似然为目标函数，并将语音识别和语音合成的概率对偶性作为约束条件，优化θ_{xy}和θ_{yx}。

可选的，基于对偶学习的语音识别与合成设备针对有标数据集中的N对有标数据$\Phi_{(x,y)}^{N}$，以最大化第一对数似然和第二对数似然为目标函数，并将语音识别和语音合成的

概率对偶性作为约束条件。理想情况下，语音识别模型和语音合成模型应满足概率对偶性，即 $P\left(x^{(i)}\right)P\left(y^{(i)}|x^{(i)},\theta_{xy}\right)=P\left(y^{(i)}\right)P\left(x^{(i)}|y^{(i)},\theta_{yx}\right)$，其中 $P\left(x^{(i)}\right)$ 和 $P\left(y^{(i)}\right)$ 分别表示语音数据 $x^{(i)}$ 和文本数据 $y^{(i)}$ 的边缘概率。目标函数和约束条件通过公式可表示如下。

$$\begin{cases} \text{目标函数 } \mathrm{F}\left(\theta_{xy},\ \theta_{yx}\right)=Max\sum_{\left(x^{(i)},y^{(i)}\right)\in\Phi_{(x,y)}{}^{N}}\left[logP\left\{f\left(x^{(i)}\right)=y^{(i)},\theta_{xy}\right\}+logP\left\{g\left(y^{(i)}\right)=x^{(i)},\theta_{yx}\right\}\right] \\ \text{约束条件 } \underset{\left(x^{(i)},y^{(i)}\right)\in\Phi_{(x,y)}{}^{N}}{st}P\left(x^{(i)}\right)P\left(y^{(i)}|x^{(i)},\theta_{xy}\right)=P\left(y^{(i)}\right)P\left(x^{(i)}|y^{(i)},\theta_{yx}\right) \end{cases}$$

可选的，联立目标函数和约束条件，采用拉格朗日乘数优化算法，将原目标函数 $\mathrm{F}\left(\theta_{xy},\ \theta_{yx}\right)$ 转化为 $\overline{\mathrm{F}}\left(\theta_{xy},\ \theta_{yx}\right)$，$\overline{\mathrm{F}}\left(\theta_{xy},\ \theta_{yx}\right)$ 表示如下。

$$\overline{\mathrm{F}}\left(\theta_{xy},\ \theta_{yx}\right)=Max\sum_{\left(x^{(i)},y^{(i)}\right)\in\Phi_{(x,y)}{}^{N}}\left[logP\left\{f\left(x^{(i)}\right)=y^{(i)},\theta_{xy}\right\}+logP\left\{g\left(y^{(i)}\right)=x^{(i)},\theta_{yx}\right\}\right.$$
$$\left.-\lambda\left(P\left(x^{(i)}\right)P\left(y^{(i)}|x^{(i)},\theta_{xy}\right)-P\left(y^{(i)}\right)P\left(x^{(i)}|y^{(i)},\theta_{yx}\right)\right)\right]$$

其中 λ 为拉格朗日因子，采用梯度下降算法，对 θ_{xy} 和 θ_{yx} 进行迭代优化。计算函数 $\overline{\mathrm{F}}\left(\theta_{xy},\ \theta_{yx}\right)$ 关于 θ_{xy} 和 θ_{yx} 的梯度，分别表示为 $\nabla_{\theta_{xy}}\overline{\mathrm{F}}\left(\theta_{xy},\ \theta_{yx}\right)$ 和 $\nabla_{\theta_{yx}}\overline{\mathrm{F}}\left(\theta_{xy},\ \theta_{yx}\right)$。对 θ_{xy} 和 θ_{xy} 进行更新，其中，$\theta_{xy}\leftarrow\left(\theta_{xy},\nabla_{\theta_{xy}}\overline{\mathrm{F}}\left(\theta_{xy},\ \theta_{yx}\right)\right)$，$\theta_{yx}\leftarrow\left(\theta_{yx},\nabla_{\theta_{yx}}\overline{\mathrm{F}}\left(\theta_{xy},\ \theta_{yx}\right)\right)$，迭代更新直到目标函数收敛或达到指定停止条件。

本申请实施例中，通过获取语音数据 $x^{(i)}$ 对应的音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率，生成文本数据 $\overline{y}^{(i)}$，获取 $y^{(i)}$ 对应的声音特征序列，生成语音数据 $\overline{x}^{(i)}$，针对 N 对有标数据，以最大化文本数据 $\overline{y}^{(i)}$ 等于文本数据 $y^{(i)}$ 的对数似然和语音数据 $\overline{x}^{(i)}$ 等于语音数据 $x^{(i)}$ 的对数似然为目标，并将语音识别和语音合成的概率对偶性作为约束条件，从而优化语音识别和语音合成效果。有效的利用了对偶学习进行语音识别与语音合成，提高语音识别和语音生成的训练速度，以及提高语音识别和语音生成输出结果的精度。

本申请实施例还提供了一种基于对偶学习的语音识别与语音合成装置，该装置能上述基于对偶学习的语音识别与语音合成方法所具备的有益效果。其中，该装置的功能可以通过硬件实现，也可以通过硬件执行相应的软件实现。所述硬件或软件包括至少一个与上述功能相对应的模块。

请参阅图 3，图 3 是本申请实施例提供的一种基于对偶学习的语音识别与语音合成装置 300 的结构框图，所述装置包括：初始化单元 301、选取单元 302、处理单元 303、第一生成单元 304、第二生成单元 305 和优化单元 306。

初始化单元 301，用于初始化有标数据集 $\Phi_{(x,y)}$、语音识别参数 θ_{xy}、语音合成参数 θ_{yx} 和

训练数据规模N，其中，有标数据集$\Phi_{(x,y)}=\left\{\left(x^{(j)},y^{(j)}\right)\right\}^{K}$，有标数据集$\Phi_{(x,y)}$中包含K对有标数据，$\left(x^{(j)},y^{(j)}\right)$表示有标数据集$\Phi_{(x,y)}$中的第j对有标数据，$x^{(j)}$为第j对有标数据中的语音数据，$y^{(j)}$为第j对有标数据中的文本数据。

选取单元302，用于从有标数据集$\Phi_{(x,y)}$中选取N对有标数据$\left\{\left(x^{(i)},y^{(i)}\right)\right\}^{N}$，K为正整数，N为小于等于K的正整数。

处理单元303，用于提取$x^{(i)}$的声学特征，根据$x^{(i)}$的声学特征，获取$x^{(i)}$对应的音素的后验概率和$x^{(i)}$对应的音素的转移概率。

第一生成单元304，用于根据$x^{(i)}$对应的音素的后验概率和$x^{(i)}$对应的音素的转移概率，生成文本数据$\overline{y}^{(i)}$，计算$\overline{y}^{(i)}$等于$y^{(i)}$的第一对数似然。

第二生成单元305，用于获取$y^{(i)}$对应的声音特征序列，并根据声音序列特征，生成语音数据$\overline{x}^{(i)}$，计算$\overline{x}^{(i)}$等于$x^{(i)}$的第二对数似然。

优化单元306，用于针对N对有标数据，以最大化第一对数似然和第二对数似然为目标函数，并将语音识别和语音合成的概率对偶性作约束条件，优化θ_{xy}和θ_{yx}。

可选的，选取单元302从有标数据集$\Phi_{(x,y)}$中选取N对有标数据$\left\{\left(x^{(i)},y^{(i)}\right)\right\}^{N}$之前，还包括：预训练单元，用于从有标数据集$\Phi_{(x,y)}$中随机选取S对有标数据，对第一语音识别模型进行预训练，得到第二语音识别模型，以及对第一语音合成模型进行预训练，得到第二语音合成模型，第二语音识别模型包括深度神经网络和隐形马尔科夫模型，第二语音合成模型包括编码器、解码器和神经声码器。

可选的，预训练单元用于对待训练的第一语音识别模型进行预训练，得到经过预训练的第二语音识别模型，具体包括：

输入单元，用于将上述S对有标数据$\left\{\left(x^{(r)},y^{(r)}\right)\right\}^{S}$中的语音数据$x^{(r)}$输入待训练的第一语音识别模型；

预处理单元，用于对语音数据$x^{(r)}$进行预处理，获取语音数据$x^{(r)}$对应的频率倒谱系数特征；

预训练单元，还用于利用频率倒谱系数特征对第一语音识别模型的深度神经网络-高斯混合模型进行预训练，得到第二语音识别模型，第二语音识别模型包括训练后的深度神经网络-高斯混合模型。

可选的，预训练单元用于对待训练的第一语音合成模型进行预训练，得到经过预训练的第二语音合成模型，具体包括：

输入单元，还用于将上述 S 对有标数据$\left\{\left(x^{(r)},y^{(r)}\right)\right\}^{S}$中的文本数据$y^{(r)}$输入待训练的第一语音合成模型；

预训练单元，还用于利用文本数据$y^{(r)}$对第一语音合成模型的编码器、解码器、神经声码器进行预训练，得到第二语言合成模型，第二语言合成模型包括训练后的编码器、训练后的解码器和训练后的神经声码器。

可选的，编码器、解码器和神经声码器均采用循环神经网络模型。

可选的，处理单元 303，包括：提取单元和获取单元。

提取单元，用于将$x^{(i)}$的声学特征输入第二语音识别模型，逐帧提取$x^{(i)}$的声学特征。

获取单元，用于将$x^{(i)}$的声学特征输入第二语音识别模型中的深层神经网络，获取$x^{(i)}$对应的音素的后验概率，通过第二语音识别模型中的隐马尔科夫模型获取$x^{(i)}$对应的音素的转移概率。

可选的，第一生成单元 304 根据语音数据$x^{(i)}$对应的音素的后验概率和语音数据$x^{(i)}$对应的音素的转移概率，生成文本数据$\overline{y}^{(i)}$，具体包括：

确定单元，用于根据语音数据$x^{(i)}$对应的音素的后验概率和语音数据$x^{(i)}$对应的音素的转移概率确定 w 条网络路径的概率，w 为大于零的正整数；

获取单元，还用于获取上述 w 条网络路径中概率最大的网络路径对应的文本数据$\overline{y}^{(i)}$。

可选的，第二生成单元 305 具体用于：将$y^{(i)}$输入第二语音合成模型的编码器，生成语义序列$\overline{\overline{y}}^{(i)}$，将$\overline{\overline{y}}^{(i)}$输入第二语音合成模型的解码器，生成声音特征序列，将声音序列特征输入第二语音合成模型的神经声码器，生成语音数据$\overline{x}^{(i)}$，计算$\overline{x}^{(i)}$等于$x^{(i)}$的第二对数似然。

可选的，优化单元 306 具体用于：以最大化第一对数似然和第二对数似然为目标函数，并将语音识别和语音合成的概率对偶性作为约束条件，联立目标函数和约束条件，采用拉格朗日乘数优化算法，对θ_{xy}和θ_{yx}进行迭代优化。

结合本申请实施例公开内容所描述的方法或者算法的步骤可以硬件的方式来实现，也可以是由处理器执行软件指令的方式来实现。软件指令可以由相应的软件模块组成，软件模块可以被存放于随机存取存储器(英文：random access memory，简称：RAM)、闪存、只读存储器(英文：read only memory，简称：ROM)、可擦除可编程只读存储器(英文：erasable programmable rom，简称：EPROM)、电可擦可编程只读存储器(英文：electrically eprom，简称：EEPROM)、寄存器、硬盘、移动硬盘、只读光盘(CD-ROM)或者本领域熟知的任何其它形式的存储介质中。一种示例性的存储介质耦合至处理器，从而使处理器能够从该存储介质读取信息，且可向该存储介质写入信息。当然，存储介质也可以是处理器的组成部分。处理器和存储介质可以位于 ASIC 中。另外，该 ASIC 可以位于网络设备中。当然，处

理器和存储介质也可以作为分立组件存在于网络设备中。

本领域技术人员应该可以意识到，在上述一个或多个示例中，本申请实施例所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机非易失性可读介质中或者作为计算机非易失性可读介质上的一个或多个指令或代码进行传输。计算机非易失性可读介质包括计算机非易失性存储介质和通信介质，其中通信介质包括便于从一个地方向另一个地方传送计算机程序的任何介质。存储介质可以是通用或专用计算机能够存取的任何可用介质。

以上所述的具体实施方式，对本申请实施例的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本申请实施例的具体实施方式而已，并不用于限定本申请实施例的保护范围，凡在本申请实施例的技术方案的基础之上，所做的任何修改、等同替换、改进等，均应包括在本申请实施例的保护范围之内。

权 利 要 求

1、一种基于对偶学习的语音识别与语音合成方法，其特征在于，所述方法包括：

初始化有标数据集$\Phi_{(x,y)}$、语音识别参数θ_{xy}、语音合成参数θ_{yx}和训练数据规模 N，其中，所述有标数据集$\Phi_{(x,y)}=\left\{\left(x^{(j)},y^{(j)}\right)\right\}^{K}$，所述有标数据集$\Phi_{(x,y)}$中包含 K 对有标数据，$\left(x^{(j)},y^{(j)}\right)$表示所述有标数据集$\Phi_{(x,y)}$中的第 j 对有标数据，$x^{(j)}$为所述第 j 对有标数据中的语音数据，$y^{(j)}$为所述第 j 对有标数据中的为文本数据，K 为正整数，N 为小于等于 K 的正整数；

从所述有标数据集$\Phi_{(x,y)}$中选取 N 对有标数据$\left\{\left(x^{(i)},y^{(i)}\right)\right\}^{N}$；

提取所述语音数据$x^{(i)}$的声学特征，根据所述语音数据$x^{(i)}$的声学特征，获取所述语音数据$x^{(i)}$对应的音素的后验概率和所述语音数据$x^{(i)}$对应的音素的转移概率；

根据所述语音数据$x^{(i)}$对应的音素的后验概率和所述语音数据$x^{(i)}$对应的音素的转移概率，生成文本数据$\bar{y}^{(i)}$，计算所述文本数据$\bar{y}^{(i)}$等于所述文本数据$y^{(i)}$的第一对数似然；

获取所述文本数据$y^{(i)}$对应的声音特征序列，并根据所述声音序列特征，生成语音数据$\bar{x}^{(i)}$，计算所述语音数据$\bar{x}^{(i)}$等于所述语音数据$x^{(i)}$的第二对数似然；

针对所述 N 对有标数据，以最大化所述第一对数似然和所述第二对数似然为目标函数，并将语音识别和语音合成的概率对偶性作为约束条件，优化所述θ_{xy}和所述θ_{yx}。

2、根据权利要求 1 所述的方法，其特征在于，所述从有标数据集$\Phi_{(x,y)}$中随机选取 N 对有标数据$\left(x^{(i)},y^{(i)}\right)$之前，所述方法还包括：

从有标数据集$\Phi_{(x,y)}$中随机选取 S 对有标数据，对待训练的第一语音识别模型进行预训练，得到经过预训练的第二语音识别模型，以及对待训练的第一语音合成模型进行预训练，得到经过预训练的第二语音合成模型，所述第二语音识别模型包括深度神经网络-隐形马尔科夫模型，所述第二语音合成模型包括编码器、解码器和神经声码器，S 为小于等于 K 的正整数。

3、根据权利要求 2 所述的方法，其特征在于，所述对待训练的第一语音识别模型进行预训练，得到经过预训练的第二语音识别模型，包括：将所述 S 对有标数据$\left\{\left(x^{(r)},y^{(r)}\right)\right\}^{S}$中的语音数据$x^{(r)}$输入待训练的所述第一语音识别模型；

对所述语音数据$x^{(r)}$进行预处理，获取所述语音数据$x^{(r)}$对应的频率倒谱系数特征；

利用所述频率倒谱系数特征对所述第一语音识别模型的深度神经网络-高斯混合模型进行预训练，得到所述第二语音识别模型，所述第二语音识别模型包括训练后的所述深度神经网络-高斯混合模型。

4、根据权利要求 2 或 3 所述的方法，其特征在于，所述对待训练的第一语音合成模型进行预训练，得到经过预训练的第二语音合成模型，包括：

将所述 S 对有标数据 $\left\{\left(x^{(r)}, y^{(r)}\right)\right\}^{S}$ 中的文本数据 $y^{(r)}$ 输入待训练的所述第一语音合成模型；

利用所述文本数据 $y^{(r)}$ 对所述第一语音合成模型的编码器、解码器、神经声码器进行预训练，得到第二语言合成模型，所述第二语言合成模型包括训练后的所述编码器、训练后的所述解码器和训练后的所述神经声码器。

5、根据权利要求 2 至 4 任一项所述的方法，其特征在于，所述编码器、所述解码器和所述神经声码器均采用循环神经网络模型。

6、根据权利要求 2 至 5 任一项所述的方法，其特征在于，所述提取所述语音数据 $x^{(i)}$ 的声学特征，获取所述语音数据 $x^{(i)}$ 对应的音素的后验概率和所述语音数据 $x^{(i)}$ 对应的音素的转移概率，包括：

将所述语音数据 $x^{(i)}$ 的声学特征输入所述第二语音识别模型，逐帧提取所述语音数据 $x^{(i)}$ 的声学特征，将所述语音数据 $x^{(i)}$ 的声学特征输入所述第二语音识别模型中的深层神经网络，获取所述语音数据 $x^{(i)}$ 对应的音素的后验概率，将所述语音数据 $x^{(i)}$ 对应的音素输入所述第二语音识别模型中的隐马尔科夫模型，获取所述语音数据 $x^{(i)}$ 对应的音素的转移概率。

7、根据权利要求 2 至 6 任一项所述的方法，其特征在于，所述根据所述语音数据 $x^{(i)}$ 对应的音素的后验概率和所述语音数据 $x^{(i)}$ 对应的音素的转移概率，生成文本数据 $\bar{y}^{(i)}$，包括：

根据所述语音数据 $x^{(i)}$ 对应的音素的后验概率和所述语音数据 $x^{(i)}$ 对应的音素的转移概率确定 w 条网络路径的概率，w 为大于零的正整数；

获取所述 w 条网络路径中概率最大的网络路径对应的文本数据 $\bar{y}^{(i)}$。

8、根据权利要求 2 至 7 任一项所述的方法，其特征在于，所述获取所述文本数据 $y^{(i)}$ 对应的声音特征序列，并根据所述声音序列特征，生成语音数据 $\bar{x}^{(i)}$，包括：

将所述文本数据 $y^{(i)}$ 输入所述第二语音合成模型的编码器，生成语义序列 $\bar{\bar{y}}^{(i)}$，将所述语义序列 $\bar{\bar{y}}^{(i)}$ 输入所述第二语音合成模型的解码器，生成声音特征序列，将所述声音序列特

征输入所述第二语音合成模型的神经声码器，生成语音数据$\bar{x}^{(i)}$。

9、根据权利要求 2 至 8 任一项所述的方法，其特征在于，所述以最大化所述第一对数似然和所述第二对数似然为目标函数，并将语音识别和语音合成的概率对偶性作为约束条件，优化所述θ_{xy}和所述θ_{yx}，包括：

以最大化所述第一对数似然和所述第二对数似然为目标函数，并将所述语音识别和语音合成的概率对偶性作为约束条件，联立所述目标函数和所述约束条件，采用拉格朗日乘数优化算法，对所述θ_{xy}和所述θ_{yx}进行迭代优化。

10、一种基于对偶学习的语音识别与语音合成装置，其特征在于，所述装置包括：

初始化单元，用于初始化有标数据集$\Phi_{(x,y)}$、语音识别参数θ_{xy}、语音合成参数θ_{yx}和训练数据规模 N，其中，所述有标数据集$\Phi_{(x,y)}=\left\{\left(x^{(j)},y^{(j)}\right)\right\}^{K}$，$\left(x^{(j)},y^{(j)}\right)$表示所述有标数据集$\Phi_{(x,y)}$中的第 j 对有标数据，所述有标数据集$\Phi_{(x,y)}$中包含 K 对有标数据，$x^{(j)}$为所述第 j 对有标数据中的语音数据，$y^{(j)}$为所述第 j 对有标数据中的文本数据，K 为正整数，N 为小于等于 K 的正整数；

选取单元，用于从所述有标数据集$\Phi_{(x,y)}$中选取 N 对有标数据$\left\{\left(x^{(i)},y^{(i)}\right)\right\}^{N}$；

处理单元，用于提取所述语音数据$x^{(i)}$的声学特征，根据所述语音数据$x^{(i)}$的声学特征，获取所述语音数据$x^{(i)}$对应的音素的后验概率和所述语音数据$x^{(i)}$对应的音素的转移概率；

第一生成单元，用于根据所述语音数据$x^{(i)}$对应的音素的后验概率和所述语音数据$x^{(i)}$对应的音素的转移概率，生成文本数据$\bar{y}^{(i)}$，计算所述文本数据$\bar{y}^{(i)}$等于所述文本数据$y^{(i)}$的第一对数似然；

第二生成单元，用于获取所述文本数据$y^{(i)}$对应的声音特征序列，并根据所述声音序列特征，生成语音数据$\bar{x}^{(i)}$，计算所述语音数据$\bar{x}^{(i)}$等于所述语音数据$x^{(i)}$的第二对数似然；

优化单元，用于针对所述 N 对有标数据，以最大化所述第一对数似然和所述第二对数似然为目标函数，并将语音识别和语音合成的概率对偶性作约束条件，优化所述θ_{xy}和所述θ_{yx}。

11、根据权利要求 10 所述的装置，其特征在于，所述装置在选取单元从所述有标数据集$\Phi_{(x,y)}$中选取 N 对有标数据$\left\{\left(x^{(i)},y^{(i)}\right)\right\}^{N}$之前，还包括：

预训练单元，用于从有标数据集$\Phi_{(x,y)}$中随机选取 S 对有标数据，对待训练的第一语音

识别模型进行预训练，得到经过预训练的第二语音识别模型，以及对待训练的第一语音合成模型进行预训练，得到经过预训练的第二语音合成模型，所述第二语音识别模型包括深度神经网络和隐形马尔科夫模型，所述第二语音合成模型包括编码器、解码器和神经声码器，S 为小于等于 K 的正整数。

12、根据权利要求 11 所述的装置，其特征在于，所述预训练单元用于对待训练的第一语音识别模型进行预训练，得到经过预训练的第二语音识别模型，具体包括：

输入单元，用于将所述 S 对有标数据 $\left\{\left(x^{(r)}, y^{(r)}\right)\right\}^{S}$ 中的语音数据 $x^{(r)}$ 输入待训练的所述第一语音识别模型；

预处理单元，用于对所述语音数据 $x^{(r)}$ 进行预处理，获取所述语音数据 $x^{(r)}$ 对应的频率倒谱系数特征；

预训练单元，还用于利用所述频率倒谱系数特征对所述第一语音识别模型的深度神经网络-高斯混合模型进行预训练，得到所述第二语音识别模型，所述第二语音识别模型包括训练后的所述深度神经网络-高斯混合模型。

13、根据权利要求 11 或 12 所述的装置，其特征在于，所述预训练单元用于对待训练的第一语音合成模型进行预训练，得到经过预训练的第二语音合成模型，具体包括：

输入单元，还用于将所述 S 对有标数据 $\left\{\left(x^{(r)}, y^{(r)}\right)\right\}^{S}$ 中的文本数据 $y^{(r)}$ 输入待训练的所述第一语音合成模型；

预训练单元，还用于利用所述文本数据 $y^{(r)}$ 对所述第一语音合成模型的编码器、解码器、神经声码器进行预训练，得到第二语言合成模型，所述第二语言合成模型包括训练后的所述编码器、训练后的所述解码器和训练后的所述神经声码器。

14、根据权利要求 11 至 13 所述的装置，其特征在于，所述编码器、所述解码器和所述神经声码器均采用循环神经网络模型。

15、根据权利要求 11 至 14 任一项所述的装置，其特征在于，所述处理单元，包括：

提取单元，用于将所述语音数据 $x^{(i)}$ 的声学特征输入所述第二语音识别模型，逐帧提取所述语音数据 $x^{(i)}$ 的声学特征；

获取单元，用于将所述语音数据 $x^{(i)}$ 的声学特征输入所述第二语音识别模型中的深层神经网络，获取所述语音数据 $x^{(i)}$ 对应的音素的后验概率，将所述语音数据 $x^{(i)}$ 对应的音素输入所述第二语音识别模型中的隐马尔科夫模型，获取所述语音数据 $x^{(i)}$ 对应的音素的转移概率。

16、根据权利要求 11 至 15 任一项所述的装置，其特征在于，所述第一生成单元根据所述语音数据 $x^{(i)}$ 对应的音素的后验概率和所述语音数据 $x^{(i)}$ 对应的音素的转移概率，生成文本数据 $\overline{y}^{(i)}$，具体包括：

确定单元，用于根据所述语音数据 $x^{(i)}$ 对应的音素的后验概率和所述语音数据 $x^{(i)}$ 对应

的音素的转移概率确定 w 条网络路径的概率，w 为大于零的正整数；

获取单元，还用于获取所述 w 条网络路径中概率最大的网络路径对应的文本数据 $\bar{y}^{(i)}$。

17、根据权利要求 11 至 16 任一项所述的装置，其特征在于，所述第二生成单元获取所述文本数据 $y^{(i)}$ 对应的声音特征序列，并根据所述声音序列特征，生成语音数据 $\bar{x}^{(i)}$，具体用于：

将所述文本数据 $y^{(i)}$ 输入所述第二语音合成模型的编码器，生成语义序列 $\bar{\bar{y}}^{(i)}$，将所述语义序列 $\bar{\bar{y}}^{(i)}$ 输入所述第二语音合成模型的解码器，生成声音特征序列，将所述声音序列特征输入所述第二语音合成模型的神经声码器，生成语音数据 $\bar{x}^{(i)}$。

18、根据权利要求 11 至 17 任一项所述的装置，其特征在于，所述优化单元具体用于：

以最大化所述第一对数似然和所述第二对数似然为目标函数，并将所述语音识别和语音合成的概率对偶性作为约束条件，联立所述目标函数和所述约束条件，采用拉格朗日乘数优化算法，对所述 θ_{xy} 和所述 θ_{yx} 进行迭代优化。

19、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上存储有计算机程序，该程序被处理器执行时实现权利要求 1 至 9 任一项所述的基于对偶学习的语音识别与语音合成方法。

20、一种服务器，其特征在于，包括：一个或多个处理器；存储器；一个或多个应用程序，其中所述一个或多个应用程序被存储在所述存储器中并被配置为由所述一个或多个处理器执行，所述一个或多个应用程序配置用于执行权利要求 1 至 9 任一项所述的基于对偶学习的语音识别与语音合成方法。

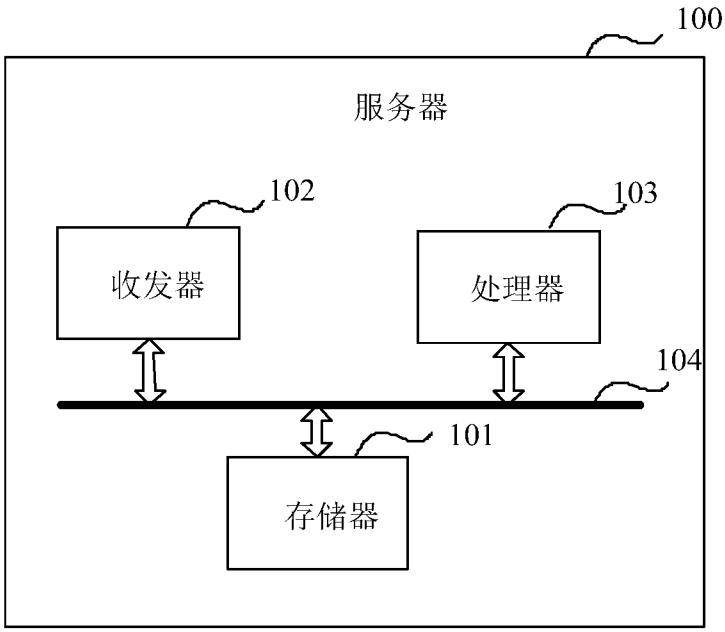

图 1

S201、初始化有标数据集$\Phi_{(x,y)}$、语音识别参数θ_{xy}、语音合成参数 θ_{yx}和训练数据规模N，其中，$\Phi_{(x,y)}=\left\{\left(x^{(j)},y^{(j)}\right)\right\}^{K}$，有标数据集 $\Phi_{(x,y)}$中包含K对有标数据， $x^{(j)}$为语音数据，$y^{(j)}$为文本数据

S202、基于对偶学习的语音识别与合成设备从有标数据集$\Phi_{(x,y)}$中选取N对有标数据，构成有标数据集$\Phi_{(x,y)}{}^{N}$，$\Phi_{(x,y)}{}^{N}=\left\{\left(x^{(i)},y^{(i)}\right)\right\}^{N}$

S203、基于对偶学习的语音识别与合成设备提取语音数据 $x^{(i)}$ 的声学特征，根据语音数据的声学特征，获取语音数据 $x^{(i)}$ 对应的音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率

S204、基于对偶学习的语音识别与合成设备根据语音数据 $x^{(i)}$ 对应的音素的后验概率和语音数据 $x^{(i)}$ 对应的音素的转移概率，生成文本数据 $\bar{y}^{(i)}$，计算文本数据 $\bar{y}^{(i)}$等于文本数据 $y^{(i)}$ 的第一对数似然

S205、基于对偶学习的语音识别与合成设备获取 $y^{(i)}$ 对应的声音特征序列特征，并根据声音序列特征，生成语音数据 $\bar{x}^{(i)}$ ，计算语音数据 $\bar{x}^{(i)}$ 等于语音数据 $x^{(i)}$ 的第二对数似然

S206、针对N对有标数据，以最大化第一对数似然和第二对数似然为目标函数，并将语音识别和语音合成的概率对偶性作为约束条件，优化 θ_{xy} 和 θ_{yx}

图 2

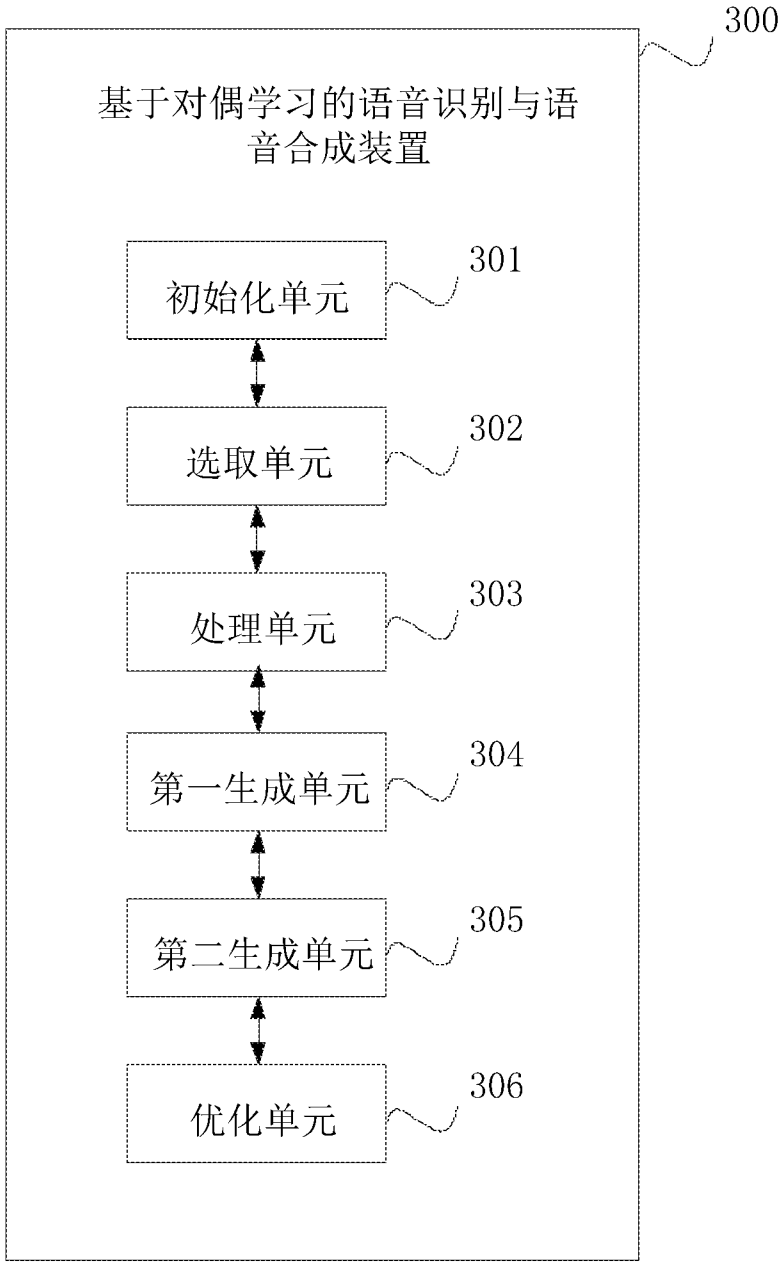
300
基于对偶学习的语音识别与语音合成装置
初始化单元
301
选取单元
302
处理单元
303
第一生成单元
304
第二生成单元
305
优化单元
306

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/117567

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

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)

CNPAT, EPODOC, WPI, CNKI, IEEE: 对偶, 学习, 合成, 语音, 识别, 概率, 训练, 数据, 模型, dual, learning, synthesis, voice, recognition, probability, train, data, model

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109887484 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 14 June 2019 (2019-06-14) claims 1-10, and description, paragraphs [0056]-[0068]	1-20
A	CN 108133705 A (RU'AN TECHNOLOGY CO., LTD.) 08 June 2018 (2018-06-08) description, paragraphs [0023]-[0044]	1-20
A	CN 108847249 A (SUZHOU AISPEECH INFORMATION TECHNOLOGY CO., LTD. et al.) 20 November 2018 (2018-11-20) entire document	1-20
A	CN 105810191 A (JIANGSU VOCATIONAL COLLEGE OF INFORMATION TECHNOLOGY) 27 July 2016 (2016-07-27) entire document	1-20
A	US 2011218804 A1 (KABUSHIKI KAISHA TOSHIBA) 08 September 2011 (2011-09-08) 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	Date of mailing of the international search report
10 February 2020	**18 February 2020**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/117567

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109887484 A	14 June 2019	None	
CN 108133705 A	08 June 2018	None	
CN 108847249 A	20 November 2018	None	
CN 105810191 A	27 July 2016	None	
US 2011218804 A1	08 September 2011	JP 2011180596 A	15 September 2011
		GB 2478314 A	07 September 2011

国际检索报告

国际申请号
PCT/CN2019/117567

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

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））

CNPAT, EPODOC, WPI, CNKI, IEEE: 对偶，学习，合成，语音，识别，概率，训练，数据，模型，dual, learning, synthesis, voice, recognition, probability, train, data, model

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	CN 109887484 A (平安科技深圳有限公司) 2019年 6月 14日 (2019-06-14) 权利要求1-10，说明书第[0056]-[0068]段	1-20
A	CN 108133705 A (儒安科技有限公司) 2018年 6月 8日 (2018-06-08) 说明书第[0023]-[0044]段	1-20
A	CN 108847249 A (苏州思必驰信息科技有限公司 等) 2018年 11月 20日 (2018-11-20) 全文	1-20
A	CN 105810191 A (江苏信息职业技术学院) 2016年 7月 27日 (2016-07-27) 全文	1-20
A	US 2011218804 A1 (KABUSHIKI KAISHA TOSHIBA) 2011年 9月 8日 (2011-09-08) 全文	1-20

☐ 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 2月 10日	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 王莹 电话号码 86-(10)-53961411

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/117567

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 109887484 A	2019年 6月 14日	无	
CN 108133705 A	2018年 6月 8日	无	
CN 108847249 A	2018年 11月 20日	无	
CN 105810191 A	2016年 7月 27日	无	
US 2011218804 A1	2011年 9月 8日	JP 2011180596 A	2011年 9月 15日
		GB 2478314 A	2011年 9月 7日