(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112802461 A
(43)申请公布日 2021.05.14

(21)申请号 202011607655.7

(22)申请日 2020.12.30

(71)申请人 深圳追一科技有限公司
地址 518051 广东省深圳市南山区粤海街道科技园社区科苑路8号讯美科技广场3号楼23A、23B

(72)发明人 周维聪 袁丁 赵金昊 刘云峰

(74)专利代理机构 广州华进联合专利商标代理有限公司 44224
代理人 方高明

(51)Int.Cl.
G10L 15/14(2006.01)
G10L 25/24(2013.01)
G10L 15/16(2006.01)
G10L 15/26(2006.01)
G10L 15/02(2006.01)

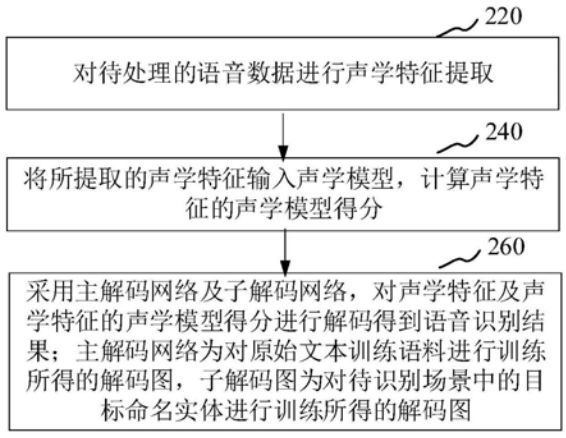
权利要求书2页 说明书11页 附图6页

(54)发明名称

语音识别方法和装置、服务器、计算机可读存储介质

(57)摘要

本申请涉及一种语音识别方法和装置、服务器、计算机可读存储介质，包括：对待处理的语音数据进行声学特征提取，将所提取的声学特征输入声学模型，计算声学特征的声学模型得分。采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码得到语音识别结果。该语音识别方法，并未对待识别场景重新训练解码网络，而是对待识别场景中的目标命名实体进行训练得到子解码图，再采用主解码网络及子解码网络进行解码得到语音识别结果。所以，针对待识别场景中的目标命名实体，基于子解码网络就可以对目标命名实体进行准确地解码。且因为未对待识别场景重新训练解码网络，所以大大缩短训练时间长，提高语音识别效率。



CN 112802461 A

1.一种语音识别方法,其特征在于,包括:

对待处理的语音数据进行声学特征提取;

将所提取的声学特征输入声学模型,计算所述声学特征的声学模型得分;

采用主解码网络及子解码网络,对所述声学特征及所述声学特征的声学模型得分进行解码得到语音识别结果;所述主解码网络为对原始文本训练语料进行训练所得的解码图,所述子解码图为对待识别场景中的目标命名实体进行训练所得的解码图。

2.根据权利要求1所述的方法,其特征在于,所述主解码网络的生成过程,包括:

对所述原始文本训练语料中的命名实体进行挖空处理,得到目标文本训练语料;

对所述目标文本训练语料进行训练得到语言模型;

对与所述原始文本训练语料对应的语音训练语料进行训练,得到声学模型;

将所述语言模型与所述声学模型进行结合得到主解码网络,所述主解码网络中包括空节点,所述空节点对应于所述目标文本训练语料中的挖空位置。

3.根据权利要求2所述的方法,其特征在于,所述将所述语言模型与所述声学模型进行结合得到主解码网络,包括:

采用compose算法将所述语言模型与所述声学模型进行结合得到主解码网络。

4.根据权利要求2所述的方法,其特征在于,所述子解码网络的生成过程,包括:

采集待识别场景中的目标命名实体构成目标命名实体文本;

对所述目标命名实体文本赋予语言模型得分;

将赋予语言模型得分的所述目标命名实体文本与所述声学模型进行结合得到子解码网络。

5.根据权利要求1所述的方法,其特征在于,所述采用主解码网络及子解码网络,对所述声学特征及所述声学特征的声学模型得分进行解码得到语音识别结果,包括:

采用所述主解码网络及所述子解码网络,对所述声学特征及所述声学特征的声学模型得分进行解码,得到语音识别网格lattice;

基于所述语音识别网格lattice得到语音识别结果。

6.根据权利要求5所述的方法,其特征在于,所述语音识别网格lattice包括多条词序列,所述词序列包括节点与跳转边,所述跳转边携带了所述声学特征的词信息;

所述采用主解码网络及子解码网络,对所述声学特征及所述声学特征的声学模型得分进行解码,得到语音识别网格lattice,包括:

从所述主解码网络中依次获取所述语音数据中各所述声学特征对应的词序列;

若所述词序列的中间节点的跳转边上的词信息为空,则调用所述子解码网络,从所述子解码网络中获取所述音频信号中下一个声学特征对应的词序列;

直到在所述子解码网络中到达所述词序列的终止节点,则返回所述主编码网络,从所述主解码网络中继续获取所述音频信号中下一个声学特征对应的词序列,直到依次获取所述待处理的音频信号中所有声学特征的词序列;

将所述所有声学特征的词序列依次连接,构成语音识别网格lattice。

7.根据权利要求6所述的方法,其特征在于,所述跳转边还携带了所述声学特征的语言模型得分;所述基于所述语音识别网格lattice得到语音识别结果,包括:

基于所述声学特征的声学模型得分、所述声学特征的语言模型得分,获取所述语音识

别网格lattice中每条词序列的总分；

获取所述词序列的总分最高的词序列作为目标词序列；

获取所述目标词序列的词信息，将所述目标词序列中的词信息作为语音识别结果。

8.根据权利要求4所述的方法，其特征在于，所述待识别场景包括医学场景、图像处理场景、电竞场景中的至少一个。

9.一种语音识别装置，其特征在于，所述装置包括：

声学特征提取模块，用于对待处理的语音数据进行声学特征提取；

声学模型得分计算模块，用于将所提取的声学特征输入声学模型，计算所述声学特征的声学模型得分；

解码模块，用于采用主解码网络及子解码网络，对所述声学特征及所述声学特征的声学模型得分进行解码得到语音识别结果；所述主解码网络为对原始文本训练语料进行训练所得的解码图，所述子解码图为对待识别场景中的目标命名实体进行训练所得的解码图。

10.根据权利要求9所述的装置，其特征在于，还包括：主解码网络生成模块，用于对原始文本训练语料中的命名实体进行挖空处理，得到目标文本训练语料；对所述目标文本训练语料进行训练得到语言模型；对与所述原始文本训练语料对应的语音训练语料进行训练，得到声学模型；将所述语言模型与所述声学模型进行结合得到主解码网络，所述主解码网络中包括空节点，所述空节点对应于所述目标文本训练语料中的挖空位置。

11.一种服务器，包括存储器及处理器，所述存储器中储存有计算机程序，其特征在于，所述计算机程序被所述处理器执行时，使得所述处理器执行如权利要求1至8中任一项所述的语音识别方法的步骤。

12.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至8中任一项所述的语音识别方法的步骤。

语音识别方法和装置、服务器、计算机可读存储介质

技术领域

[0001] 本申请涉及自然语言处理技术领域，特别是涉及一种语音识别方法和装置、服务器、计算机可读存储介质。

背景技术

[0002] 随着人工智能和自然语言处理技术的不断发展，语音识别技术也得到了快速地发展。采用语音识别技术可以自动将音频信号转变为相应的文本或命令。传统的语音识别技术可以应用在普通的、日常的语音识别场景中，并取得较好的识别效果。

[0003] 但是，当应用于专业场景下时，由于专业场景下包含大量的专业词汇，所以采用传统的语音识别技术，识别效果较差。而如果专门针对该专业场景重新训练解码网络进行语音识别，显然，重新训练解码图的工作量较大、训练时间长、无法快速实现。

发明内容

[0004] 本申请实施例提供一种语音识别方法、装置、服务器、计算机可读存储介质，可以在针对特定应用场景下进行语音识别时，减小重新训练解码图的工作量、并大大缩短训练时间长，提高语音识别的效率。

[0005] 一种语音识别方法，包括：

[0006] 对待处理的语音数据进行声学特征提取；

[0007] 将所提取的声学特征输入声学模型，计算所述声学特征的声学模型得分；

[0008] 采用主解码网络及子解码网络，对所述声学特征及所述声学特征的声学模型得分进行解码得到语音识别结果；所述主解码网络为对原始文本训练语料进行训练所得的解码图，所述子解码图为对待识别场景中的目标命名实体进行训练所得的解码图。

[0009] 一种语音识别装置，所述装置包括：

[0010] 声学特征提取模块，用于对待处理的语音数据进行声学特征提取；

[0011] 声学模型得分计算模块，用于将所提取的声学特征输入声学模型，计算所述声学特征的声学模型得分；

[0012] 解码模块，用于采用主解码网络及子解码网络，对所述声学特征及所述声学特征的声学模型得分进行解码得到语音识别结果；所述主解码网络为对原始文本训练语料进行训练所得的解码图，所述子解码图为对待识别场景中的目标命名实体进行训练所得的解码图。

[0013] 一种服务器，包括存储器及处理器，所述存储器中储存有计算机程序，所述计算机程序被所述处理器执行时，使得所述处理器执行如上方法的步骤。

[0014] 一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现如上方法的步骤。

[0015] 上述语音识别方法、装置、服务器、计算机可读存储介质，对待处理的语音数据进行声学特征提取，将所提取的声学特征输入声学模型，计算声学特征的声学模型得分。采用

主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码得到语音识别结果。由于子解码网络为对待识别场景中的目标命名实体进行训练所得的解码网络。该语音识别方法，并未对待识别场景重新训练解码网络，而是对待识别场景中的目标命名实体进行训练得到子解码网络，再采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码得到语音识别结果。所以，针对待识别场景中的目标命名实体，基于子解码网络就可以对目标命名实体进行准确地解码。且因为未对待识别场景重新训练解码网络，所以大大缩短训练时间长，提高语音识别效率。

附图说明

[0016] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0017] 图1为一个实施例中语音识别方法的应用场景图；

[0018] 图2为一个实施例中语音识别方法的流程图；

[0019] 图3为一个实施例中主解码网络生成方法的流程图；

[0020] 图4为一个实施例中主解码网络的部分结构示意图；

[0021] 图5为一个实施例中子解码网络生成方法的流程图；

[0022] 图6为一个实施例中语音识别网格lattice的结构示意图；

[0023] 图7为一个实施例中采用主解码网络及子解码网络进行解码，得到语音识别网格lattice方法的流程图；

[0024] 图8为一个实施例中采用主解码网络及子解码网络进行解码，得到语音识别网格lattice的示意图；

[0025] 图9为一个实施例中基于语音识别网格lattice得到语音识别结果方法的流程图；

[0026] 图10为一个实施例中语音识别装置的结构框图；

[0027] 图11为另一个实施例中语音识别装置的结构框图；

[0028] 图12为一个实施例中服务器的内部结构示意图。

具体实施方式

[0029] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0030] 可以理解，本申请所使用的术语“第一”、“第二”等可在本文中用于描述各种元件，但这些元件不受这些术语限制。这些术语仅用于将第一个元件与另一个元件区分。

[0031] 图1为一个实施例中语音识别方法的应用场景图。如图1所示，该应用环境包括终端120和服务器140，该终端120与服务器140之间通过网络连接。服务器140通过本申请中的语音识别方法，对待处理的语音数据进行声学特征提取；将所提取的声学特征输入声学模型，计算声学特征的声学模型得分；采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码得到语音识别结果；主解码网络为对原始文本训练语料进行训练

所得的解码图，子解码图为对待识别场景中的目标命名实体进行训练所得的解码图。这里，终端120可以是手机、平板电脑、PDA(Personal Digital Assistant，个人数字助理)、车载电脑、穿戴式设备、智能家居等任意终端设备。

[0032] 图2为一个实施例中语音识别方法的流程图，如图2所示，提供了一种语音识别方法，应用于服务器，包括步骤220至步骤260。其中，

[0033] 步骤220，对待处理的语音数据进行声学特征提取。

[0034] 其中，语音数据可以是指所获取到的音频信号。具体可以是在语音输入场景、智能聊天场景、语音翻译场景中所获取到的音频信号。对待处理的语音数据进行声学特征提取。声学特征提取的具体过程可以是：将获取到的一维音频信号，通过特征提取算法转化为一组高维向量。所得到的高维向量即为声学特征，常见的声学特征有MFCC，Fbank，ivector等，本申请对此不做限定。Fbank(FilterBank)就是一种前端处理算法，以类似于人耳的方式对音频进行处理，可以提高语音识别的性能。其中，获得语音信号的Fbank特征的一般步骤是：预加重、分帧、加窗、短时傅里叶变换(STFT)、mel滤波、去均值等。且通过对Fbank做离散余弦变换(DCT)即可获得MFCC特征。

[0035] 其中，MFCC(梅尔频率倒谱系数，Mel-frequency cepstral coefficients)，梅尔频率是基于人耳听觉特性提出来的，因此，梅尔频率与Hz频率成非线性对应关系。梅尔频率倒谱系数(MFCC)则是利用梅尔频率与Hz频率之间的这种非线性对应关系，计算得到的Hz频谱特征。MFCC主要用于语音数据特征提取和降低运算维度。例如：对于一帧有512维(采样点)数据，经过MFCC后可以提取出最重要的40维数据，从而达到了降维的目的。其中，ivector就是描述每个说话人的特征向量。

[0036] 步骤240，将所提取的声学特征输入声学模型，计算声学特征的声学模型得分。

[0037] 具体的，声学模型可以包括神经网络模型和隐马尔可夫模型，其中，神经网络模型可以向隐马尔可夫模型提供声学建模单元，声学建模单元的粒度可以包括：字、音节、音素、或者状态等。而隐马尔可夫模型可以依据神经网络模型提供的声学建模单元，确定音素序列。一个状态在数学上表征一个马尔科夫过程的状态。且声学模型是预先根据音频训练语料进行训练，所得到的模型。

[0038] 将所提取的声学特征输入声学模型，可以计算得到声学特征的声学模型得分。其中，声学模型得分可以看作是根据每个声学特征下每个音素出现的概率计算得到的分数。

[0039] 步骤260，采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码得到语音识别结果；主解码网络为对原始文本训练语料进行训练所得的解码图，子解码图为对待识别场景中的目标命名实体进行训练所得的解码图。

[0040] 其中，解码网络用于在给定音素序列的情况下，找到最佳的解码路径，进而可以得到语音识别结果。在本申请实施例中，所采用的解码网络包括主解码网络及子解码网络，主解码网络为对原始文本训练语料进行训练所得的解码图，子解码图为对待识别场景中的目标命名实体进行训练所得的解码图。如此，采用主解码网络可以对除去命名实体的音素序列进行解码，采用子解码网络可以对目标命名实体的音素序列进行解码。从而，采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码，就得到了语音识别结果。

[0041] 其中，待识别场景中的目标命名实体包括通过人工在待识别场景下所指定的一些

语音识别错误率超过预设错误率阈值的命名实体，还包括待识别场景中的专业词汇，本申请对此不做限定。

[0042] 本申请实施例中，针对特定应用场景下语音识别的准确率较低的技术问题，提出了一种语音识别方法，对待处理的语音数据进行声学特征提取，将所提取的声学特征输入声学模型，计算声学特征的声学模型得分。采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码得到语音识别结果。该语音识别方法，并未对待识别场景重新训练解码网络，而是对待识别场景中的目标命名实体进行训练得到子解码图，再采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码得到语音识别结果。所以，针对待识别场景中的目标命名实体，基于子解码网络就可以对目标命名实体进行准确地解码。且因为未对待识别场景重新训练解码网络，所以大大缩短训练时间长，提高语音识别效率。

[0043] 在一个实施例中，如图3所示，主解码网络的生成过程，包括：

[0044] 步骤320，对原始文本训练语料中的命名实体进行挖空处理，得到目标文本训练语料。

[0045] 获取原始文本训练语料，可以是从语料库中去获取到原始文本训练语料。该原始文本训练语料中大概包含700-1000万个文本语料。其中，语料库中存放的是在语言的实际使用中真实出现过的语言材料，且语料库是以电子计算机为载体承载语言知识的基础资源。一般情况下，真实语料需要经过加工(例如，分析和处理)，才能成为有用的资源。命名实体(named entity)，顾名思义，命名实体就是人名、机构名、地名以及其他所有以名称为标识的实体，更广泛的实体还包括数字、日期、货币、地址等。

[0046] 因为不同的应用场景下的命名实体有较大的差别，且这些命名实体也可能不是原始文本训练语料中所包含的语料。所以，为了实现提高针对特定应用场景下的语音识别的准确率，可以先对原始文本训练语料中的命名实体进行挖空处理，挖空位置用空节点来表示。从而，留下不包含命名实体的训练语料，这些不包含命名实体的训练语料，就构成了目标文本训练语料。

[0047] 步骤340，对目标文本训练语料进行训练得到语言模型。

[0048] 在得到了目标文本训练语料之后，对目标文本训练语料进行训练得到语言模型。这里的语言模型可以是采用循环神经网络进行训练，也可称为RNNLM(Recurrent Neural Network Based LanguageModel，循环网络语言模型)。循环网络语言模型除了考虑当前输入的词语外，可以考虑之前输入的多个词语，并可根据之前输入的词语组成的长文本来计算接下来出现词的概率，因此循环网络语言模型具有“更好的记忆效应”。比如“我的”“心情”后面可能出现“不错”，也可能出现“不好”，这些词的出现，依赖于之前出现了“我的”和“心情”，这个就是“记忆效应”。

[0049] 步骤360，对与原始文本训练语料对应的语音训练语料进行训练，得到声学模型。

[0050] 然后，获取原始文本训练语料对应的语音训练语料，这里的语音训练语料与原始文本训练语料中的各语料之间存在对应关系。即原始文本训练语料中的每个文本语料，都可以在语音训练语料中获取到对应的语音语料。

[0051] 声学模型采用深度神经网络模型对声学发音和基本声学单元(通常是音素)之间的映射关系进行建模。其中，音素是根据语音的自然属性划分出来的最小语音单位。声学模

型可以接收输入的声学特征，并输出声学特征对应的音素序列。针对语音数据库中的语音语料进行声学特征提取，依据提取得到的声学特征进行声学模型的训练，从而得到声学模型。

[0052] 步骤380，将语言模型与声学模型进行结合得到主解码网络，主解码网络中包括空节点，空节点对应于目标文本训练语料中的挖空位置。

[0053] 因为是对原始文本训练语料中的命名实体进行挖空处理，得到目标文本训练语料，并对目标文本训练语料进行训练得到语言模型。且对与原始文本训练语料对应的语音训练语料进行训练，得到声学模型。所以，将语言模型与声学模型进行结合就可以得到主解码网络。该主解码网络中包含了除去命名实体之外的训练语料对应的节点，对于命名实体采用空节点来表示，空节点对应于目标文本训练语料中的挖空位置即命名实体对应的位置。

[0054] 如图4所示，为一个实施例中主解码网络的部分结构示意图。其中，一个词序列包括节点和跳转边，其中节点包括起始节点、中间节点和终止节点。如图4所示，节点1为起始节点，节点2、3、4、5为中间节点，节点6为终止节点。起始节点与终止节点之间连接了跳转边，跳转边上携带了词信息与音素信息。其中，节点1与节点2之间的跳转边携带了词信息：你好；携带了音素信息为：n。节点2与节点3之间的跳转边携带了词信息：blank；携带了音素信息为：i。节点3与节点4之间的跳转边携带了词信息：blank；携带了音素信息为：h。节点4与节点5之间的跳转边携带了词信息：blank；携带了音素信息为：ao。节点5即为将命名实体进行挖空后所得的空节点，因此，节点5与节点6之间的跳转边未携带词信息，也未携带音素信息。

[0055] 本申请实施例中，对原始文本训练语料中的命名实体进行挖空处理，得到目标文本训练语料，对目标文本训练语料进行训练得到语言模型。对与原始文本训练语料对应的语音训练语料进行训练，得到声学模型。将语言模型与声学模型进行结合得到主解码网络，主解码网络中包括空节点，空节点对应于目标文本训练语料中的挖空位置。对命名实体进行挖空后所得的主解码网络，就可以与任意特定场景下的目标命名实体训练所得的子解码网络进行组合，以适用于任意特定场景下的语音识别，提高任意特定场景下语音识别的准确率和效率。

[0056] 在一个实施例中，将语言模型与声学模型进行结合得到主解码网络，包括：

[0057] 采用compose算法将语言模型与声学模型进行结合得到主解码网络。

[0058] 本申请实施例中，通过compose算法可以实现将第一个WFST的某个转移上的输出标签等于第二个WFST的某个转移上的输入标签，然后把这些转移上的label和weight分别进行操作。compose算法的具体实现代码，在本申请中不在赘述。

[0059] 在一个实施例中，如图5所示，子解码网络的生成过程，包括：

[0060] 步骤520，采集待识别场景中的目标命名实体构成目标命名实体文本。

[0061] 其中，待识别场景中的目标命名实体包括通过人工在待识别场景下所指定的一些语音识别错误率超过预设错误率阈值的命名实体。另外，还可以将待识别场景下的专业词汇作为目标命名实体文本，例如对于医学场景，将医生、患者、血压、心跳、CT(computed tomography)等专业词汇作为目标命名实体文本。而对于电竞场景，则将吃鸡、MVP等专业词汇作为目标命名实体文本。显然，不同的应用场景下的目标命名实体之间差距很大。

[0062] 分别采集不同应用场景下的目标命名实体，构成各应用场景下的目标命名实体文本。

[0063] 步骤540，对所述目标命名实体文本赋予语言模型得分。

[0064] 这里的语言模型可以是采用循环神经网络进行训练，也可称为RNNLM(Recurrent Neural Network Based LanguageModel，循环网络语言模型)。循环网络语言模型除了考虑当前输入的词语外，可以考虑之前输入的多个词语，并可根据之前输入的词语组成的长文本来计算接下来出现词的概率，因此循环网络语言模型具有“更好的记忆效应”。比如“我的”“心情”后面可能出现“不错”，也可能出现“不好”，这些词的出现，依赖于之前出现了“我的”和“心情”，这个就是“记忆效应”。

[0065] 其中，可以通过人工对所述目标命名实体文本赋予语言模型得分，一般情况下采用为目标命名实体文本赋予较高的语言模型得分，来提高这些目标命名实体的识别准确率。此处，较高的语言模型得分可以是超过预设得分阈值的分数，例如设置预设得分阈值为0.9。

[0066] 步骤560，将赋予语言模型得分的目标命名实体文本与声学模型进行结合得到子解码网络。

[0067] 采用compose算法将赋予语言模型得分的目标命名实体文本与声学模型进行结合得到子解码网络。

[0068] 本申请实施例中，采集待识别场景中的目标命名实体构成目标命名实体文本，对所述目标命名实体文本赋予语言模型得分。再将赋予语言模型得分的目标命名实体文本与声学模型进行结合得到子解码网络。其中，子解码网络可以插入至主解码网络的空节点处，以通过主解码网络及子解码网络对语音数据进行准确完整的语音识别。

[0069] 在一个实施例中，采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码得到语音识别结果，包括：

[0070] 采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码，得到语音识别网格lattice；

[0071] 基于语音识别网格lattice得到语音识别结果。

[0072] 具体的，采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码，得到语音识别网格lattice。即采用主解码网络对除去目标命名实体之外的声学特征进行解码，采用子解码网络对目标命名实体对应的声学特征进行解码，得到语音识别网格lattice。

[0073] 语音识别网格lattice包括多条备选词序列。其中，备选词序列包括多个词语以及多个路径，也可称为lattice，lattice本质上是一个有向无环(directed acyclic graph)图，图上的每个节点代表一个词的结束时间点，每条跳转边代表一个可能的词，以及该词发生的声学模型得分和语言模型得分。对语音识别的结果进行表示时，每个节点存放当前位置上语音识别的结果，包括声学概率、语言概率等信息。

[0074] 如图6所示，为一个实施例中语音识别网格lattice的结构示意图。从最左边起始节点开始沿着不同的弧走到终止节点，可以得到不同的词序列，而各个弧上存放的概率组合起来，表示了输入语音得到某一段文字的概率(分数)。比如图6中所示的，“你好深圳”、“你好北京”、“你好背景”，都可以看作是语音识别结果的一条路径，即“你好深圳”、“你好北

京”、“你好背景”均是一个词序列，这些词序列就构成了语音识别网格lattice。且图中每条路径均对应有一个概率，根据概率可计算得到每条路径的分数。

[0075] 显然，所得到的语音识别网格lattice比较庞大，因此可以对语音识别网格lattice进行剪枝。其中，一种剪枝方法是对lattice进行前后向打分，计算每条跳转边的后验概率，然后删除后验概率较低的跳转边。从而，经过对语音识别网格lattice进行剪枝之后，就将语音识别网格lattice进行了简化，但是仍然保留了语音识别网格lattice中的重要信息。

[0076] 在基于剪枝处理之后所得的语音识别网格lattice，从其中的词序列中提取得分靠前的预设数量的词序列作为候选词序列。从候选词序列中筛选出目标词序列作为语音识别结果。此处，目标词序列的数目一般为一个。

[0077] 本申请实施例中，采用主解码网络对除去目标命名实体之外的声学特征进行解码，采用子解码网络对目标命名实体对应的声学特征进行解码，得到语音识别网格lattice。所得到的语音识别网格lattice包括多条备选词序列，因此，首先对语音识别网格lattice进行剪枝，然后再对剪枝处理之后的语音识别网格lattice进行筛选，最终筛选出目标词序列作为语音识别结果。针对待识别场景中的目标命名实体，基于子解码网络就可以对目标命名实体进行准确地解码。再对经过主解码网络和子解码网络解码后所得的语音识别网格lattice进行剪枝并筛选出目标词序列作为语音识别结果。因此，提高了语音识别效率和准确率。

[0078] 在一个实施例中，语音识别网格lattice包括多条词序列，词序列包括节点与跳转边，跳转边携带了声学特征的词信息；

[0079] 如图7所示，采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码，得到语音识别网格lattice，包括：

[0080] 步骤720，从主解码网络中依次获取语音数据中各声学特征对应的词序列。

[0081] 结合图8所示，为一个实施例中解码过程的示意图。对待处理的语音数据进行声学特征提取。将所提取的声学特征输入声学模型，计算声学特征的声学模型得分。这里的声学特征包括音素，当然本申请对此不做限定。例如，接收到一段音频信号，并从中依次提取到的声学特征为n，i，h，ao，sh，en，zh，en这八个音素，并从主解码网络中依次获取这八个音素对应的词序列。

[0082] 如图8所示，从主解码网络中依次获取这八个音素对应的词序列的过程中得到，节点1为起始节点，节点2、3、4、5为中间节点，节点6为终止节点。起始节点与终止节点之间连接了跳转边，跳转边上携带了词信息与音素信息。其中，节点1与节点2之间的跳转边携带了词信息：你好；携带了音素信息为：n。节点2与节点3之间的跳转边携带了词信息：blank；携带了音素信息为：i。节点3与节点4之间的跳转边携带了词信息：blank；携带了音素信息为：h。节点4与节点5之间的跳转边携带了词信息：blank；携带了音素信息为：ao。节点5即为将命名实体进行挖空后所得的空节点，因此，节点5与节点6之间的跳转边未携带词信息，也未携带音素信息。

[0083] 步骤740，若词序列的中间节点的跳转边上的词信息为空，则调用子解码网络，从子解码网络中获取音频信号中下一个声学特征对应的词序列。

[0084] 在从主解码网络中依次获取个声学特征对应的词序列的过程中，若词序列的中间

节点的跳转边上的词信息为空，则说明对该中间节点处的命名实体进行了挖空处理。因此，若词序列的中间节点的跳转边上的词信息为空，则调用子解码网络，从子解码网络中获取音频信号中下一个声学特征对应的词序列。

[0085] 如图8所示，节点5的的跳转边之间未携带词信息，也未携带音素信息，即节点5的跳转边上的词信息为空。此时，调用子解码网络，从子解码网络中获取音频信号中下一个声学特征对应的词序列。即从子解码网络中获取音频信号中下一个声学特征sh、并依次获取en，zh，en对应的词序列。

[0086] 步骤760，直到在子解码网络中到达词序列的终止节点，则返回主编码网络，从主解码网络中继续获取音频信号中下一个声学特征对应的词序列，直到依次获取待处理的音频信号中所有声学特征的词序列。

[0087] 直到在子解码网络中到达词序列的终止节点，可以理解为完整识别出一个词序列为止。例如，结合图8可知，在子解码网络中完整识别出中sh，en，zh，en这几个因素对应的词序列为“深圳”即为到达了词序列的终止节点。当然，也可能在子解码网络中识别出sh，en，zh，en这几个因素对应的词序列为“神针”、“申震”等多种备选词序列。此时，返回主解码网络的终止节点6。然后，从主解码网络中继续获取音频信号中下一个声学特征对应的词序列，直到依次获取待处理的音频信号中所有声学特征的词序列。例如，在“你好深圳”的音频信号之后，还有“我喜欢你”的音频信号，则从主解码网络中继续获取该音频信号中下一个声学特征wo、并依次获取x，i，h，u，an，n，i对应的词序列，直到依次获取待处理的音频信号中所有声学特征的词序列。

[0088] 步骤780，将所有声学特征的词序列依次连接，构成语音识别网格lattice。

[0089] 将对语音数据中所有声学特征的词序列依次连接，就构成了语音识别网格lattice。例如，对图8中该语音数据的词序列依次连接，可以得到“你好深圳”、“你好神针”及“你好申震”等多个备选词序列。这些备选词序列就构成了语音识别网格lattice。

[0090] 本申请实施例中，从主解码网络中依次获取语音数据中各声学特征对应的词序列。若词序列的中间节点的跳转边上的词信息为空，则调用子解码网络，从子解码网络中获取音频信号中下一个声学特征对应的词序列，直到在子解码网络中到达词序列的终止节点，则返回主编码网络，从主解码网络中继续获取音频信号中下一个声学特征对应的词序列，直到依次获取待处理的音频信号中所有声学特征的词序列。最后，将所有声学特征的词序列依次连接，构成语音识别网格lattice。

[0091] 在采用主解码网络进行解码过程中，若词序列的中间节点的跳转边的词信息为空，则调用子解码网络去进行解码。最终，基于经过主解码网络及子解码网络进行解码所得的备选词序列，得到语音识别网格lattice。其中，通过子解码网络实现了对目标命名实体的准确识别。

[0092] 在一个实施例中，如图9所示，跳转边还携带了声学特征的语言模型得分；基于语音识别网格lattice得到语音识别结果，包括：

[0093] 步骤920，基于声学特征的声学模型得分、声学特征的语言模型得分，获取语音识别网格lattice中每条词序列的总分；

[0094] 步骤940，获取词序列的总分最高的词序列作为目标词序列；

[0095] 步骤960，获取目标词序列的词信息，将目标词序列中的词信息作为语音识别结

果。

[0096] 语音识别网格lattice包括多条备选词序列。其中，词序列包括多个词语以及多个路径，也可称为lattice，lattice本质上是一个有向无环(directed acyclic graph)图，图上的每个节点代表一个词的结束时间点，每条跳转边代表一个可能的词，以及该词发生的声学模型得分和语言模型得分。

[0097] 对于每条词序列，将每条跳转边上的声学模型得分和语言模型得分进行求和运算，得到了语音识别网格lattice中每条词序列的总分。再从中获取词序列的总分最高的词序列作为目标词序列。将该目标词序列的词信息，将目标词序列中的词信息作为语音识别结果。

[0098] 结合图8所示，若“你好深圳”这条词序列的总分最高，则就将该词序列的词信息即“你好深圳”，作为语音识别结果。

[0099] 本申请实施例中，对于每条词序列，将每条跳转边上的声学模型得分和语言模型得分进行求和运算，得到了语音识别网格lattice中每条词序列的总分。再从中获取词序列的总分最高的词序列作为目标词序列。将该目标词序列的词信息，将目标词序列中的词信息作为语音识别结果。如此，通过声学模型得分及语言模型得分，就可以准确筛选出目标词序列。

[0100] 在一个实施例中，待识别场景包括医学场景、图像处理场景、电竞场景中的至少一个。

[0101] 对于医学场景，则该应用场景下的目标命名实体多为专业词汇，例如，命名实体为医生、患者、血压、心跳、CT(computed tomography)等。而对于图像处理场景，则目标命名实体为分辨率、色差、逆光等。而对于电竞场景，则目标命名实体为吃鸡、MVP等。显然，不同的应用场景下的目标命名实体之间差距很大。

[0102] 本申请实施例中，针对上述不同应用场景，包括但不限于医学场景、图像处理场景、电竞场景，分别采集不同应用场景下的目标命名实体，构成各应用场景下的目标命名实体文本。然后，对所述目标命名实体文本赋予语言模型得分。最后，将赋予语言模型得分的目标命名实体文本与声学模型进行结合得到子解码网络。从而，实现对不同应用场景下的目标命名实体都能够针对性地进行语音识别，以提高最终所得到的语音识别结果的准确性。

[0103] 在一个实施例中，如图10所示，一种语音识别装置1000，包括：

[0104] 声学特征提取模块1020，用于对待处理的语音数据进行声学特征提取；

[0105] 声学模型得分计算模块1040，用于将所提取的声学特征输入声学模型，计算声学特征的声学模型得分；

[0106] 解码模块1060，用于采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码得到语音识别结果；所述主解码网络为对原始文本训练语料进行训练所得的解码图，所述子解码图为对待识别场景中的目标命名实体进行训练所得的解码图。

[0107] 在一个实施例中，如图11所示，还提供了一种语音识别装置1000，还包括：主解码网络生成模块1070，用于对原始文本训练语料中的命名实体进行挖空处理，得到目标文本训练语料；对目标文本训练语料进行训练得到语言模型；对与原始文本训练语料对应的语音训练语料进行训练，得到声学模型；将语言模型与声学模型进行结合得到主解码网络，主

解码网络中包括空节点，空节点对应于目标文本训练语料中的挖空位置。

[0108] 在一个实施例中，主解码网络生成模块1070，还用于采用compose算法将语言模型与声学模型进行结合得到主解码网络。

[0109] 在一个实施例中，如图11所示，还提供了一种语音识别装置1000，还包括：子解码网络生成模块1080，用于采集待识别场景中的目标命名实体构成目标命名实体文本；对所述目标命名实体文本赋予语言模型得分；将赋予语言模型得分的目标命名实体文本与声学模型进行结合得到子解码网络。

[0110] 在一个实施例中，解码模块1060，还包括：

[0111] 语音识别网格生成单元，用于采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码，得到语音识别网格lattice；

[0112] 语音识别结果确定单元，用于基于语音识别网格lattice得到语音识别结果。

[0113] 在一个实施例中，语音识别网格lattice包括多条词序列，词序列包括节点与跳转边，跳转边携带了声学特征的词信息；语音识别网格生成单元，还用于从主解码网络中依次获取语音数据中各声学特征对应的词序列；若词序列的中间节点的跳转边上的词信息为空，则调用子解码网络，从子解码网络中获取音频信号中下一个声学特征对应的词序列；直到在子解码网络中到达词序列的终止节点，则返回主编码网络，从主解码网络中继续获取音频信号中下一个声学特征对应的词序列，直到依次获取待处理的音频信号中所有声学特征的词序列；将所有声学特征的词序列依次连接，构成语音识别网格lattice。

[0114] 在一个实施例中，跳转边还携带了声学特征的语言模型得分；语音识别结果确定单元，还用于基于声学特征的声学模型得分、声学特征的语言模型得分，获取语音识别网格lattice中每条词序列的总分；获取词序列的总分最高的词序列作为目标词序列；获取目标词序列的词信息，将目标词序列中的词信息作为语音识别结果。

[0115] 在一个实施例中，待识别场景包括医学场景、图像处理场景、电竞场景中的至少一个。

[0116] 上述语音识别装置中各个模块的划分仅用于举例说明，在其他实施例中，可将语音识别装置按照需要划分为不同的模块，以完成上述语音识别装置的全部或部分功能。

[0117] 图12为一个实施例中服务器的内部结构示意图。如图12所示，该服务器包括通过系统总线连接的处理器和存储器。其中，该处理器用于提供计算和控制能力，支撑整个服务器的运行。存储器可包括非易失性存储介质及内存储器。非易失性存储介质存储有操作系统和计算机程序。该计算机程序可被处理器所执行，以用于实现以下各个实施例所提供的一种语音识别方法。内存储器为非易失性存储介质中的操作系统计算机程序提供高速缓存的运行环境。该服务器可以是手机、平板电脑或者个人数字助理或穿戴式设备等。

[0118] 本申请实施例中提供的语音识别装置中的各个模块的实现可为计算机程序的形式。该计算机程序可在终端或服务器上运行。该计算机程序构成的程序模块可存储在终端或服务器的存储器上。该计算机程序被处理器执行时，实现本申请实施例中所描述方法的步骤。

[0119] 本申请实施例还提供了一种计算机可读存储介质。一个或多个包含计算机可执行指令的非易失性计算机可读存储介质，当计算机可执行指令被一个或多个处理器执行时，使得处理器执行语音识别方法的步骤。

[0120] 一种包含指令的计算机程序产品，当其在计算机上运行时，使得计算机执行语音识别方法。

[0121] 本申请实施例所使用的对存储器、存储、数据库或其它介质的任何引用可包括非易失性和/或易失性存储器。合适的非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)，它用作外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDR SDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)。

[0122] 以上实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对本申请专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

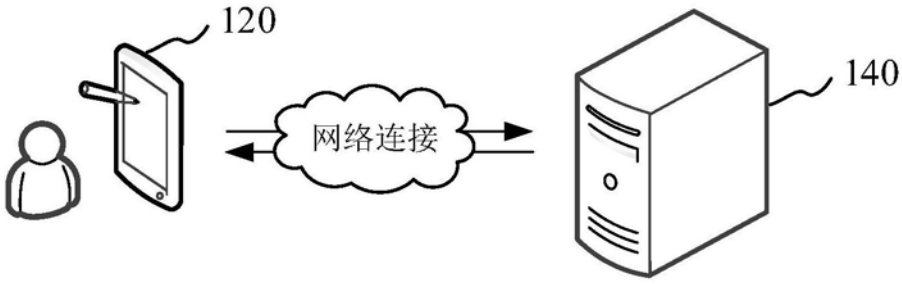

120
网络连接
140

图1

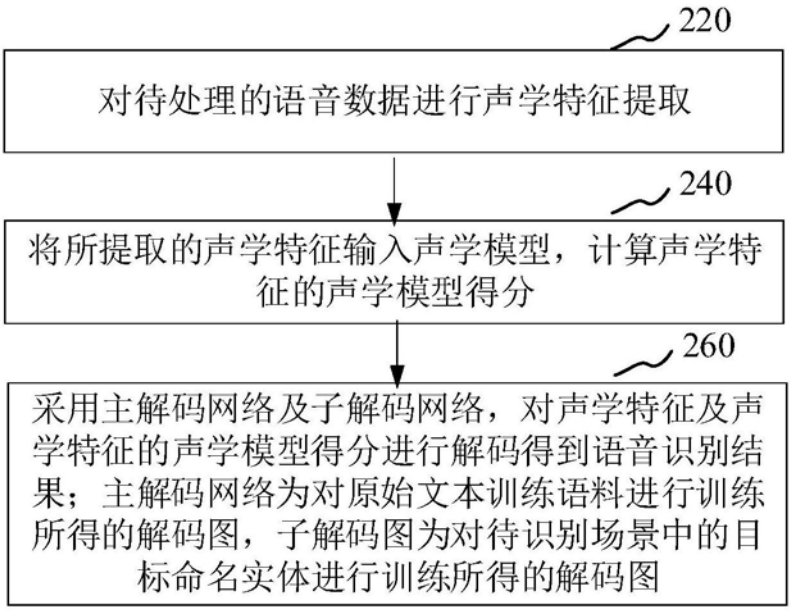

220
对待处理的语音数据进行声学特征提取
240
将所提取的声学特征输入声学模型，计算声学特征的声学模型得分
260
采用主解码网络及子解码网络，对声学特征及声学特征的声学模型得分进行解码得到语音识别结果；主解码网络为对原始文本训练语料进行训练所得的解码图，子解码图为对待识别场景中的目标命名实体进行训练所得的解码图

图2

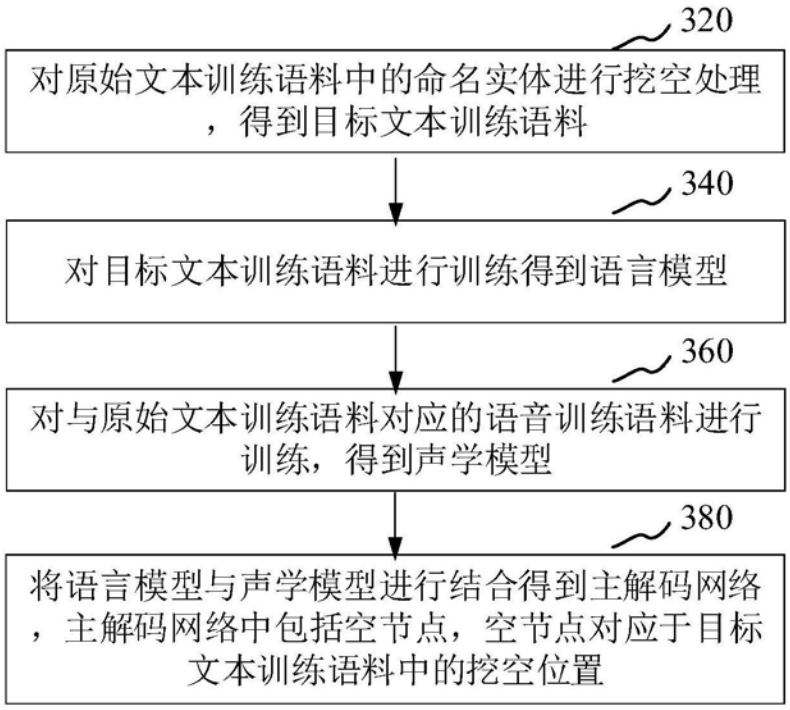

320
对原始文本训练语料中的命名实体进行挖空处理，得到目标文本训练语料
340
对目标文本训练语料进行训练得到语言模型
360
对与原始文本训练语料对应的语音训练语料进行训练，得到声学模型
380
将语言模型与声学模型进行结合得到主解码网络，主解码网络中包括空节点，空节点对应于目标文本训练语料中的挖空位置

图3

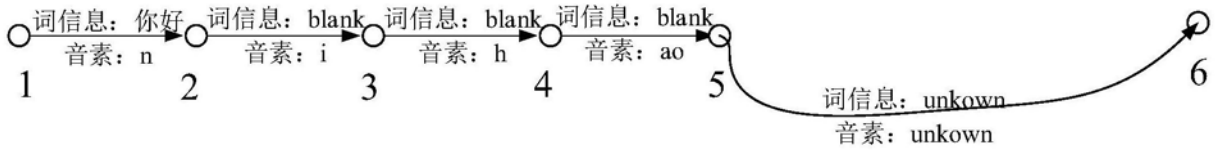

词信息：你好
音素：n
词信息：blank
音素：i
词信息：blank
音素：h
词信息：blank
音素：ao
词信息：unkown
音素：unkown
1
2
3
4
5
6

图4

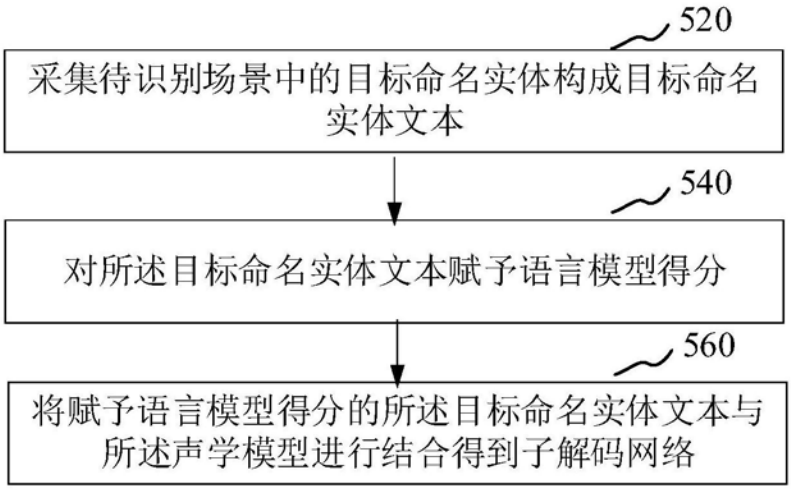

520
采集待识别场景中的目标命名实体构成目标命名实体文本
540
对所述目标命名实体文本赋予语言模型得分
560
将赋予语言模型得分的所述目标命名实体文本与所述声学模型进行结合得到子解码网络

图5

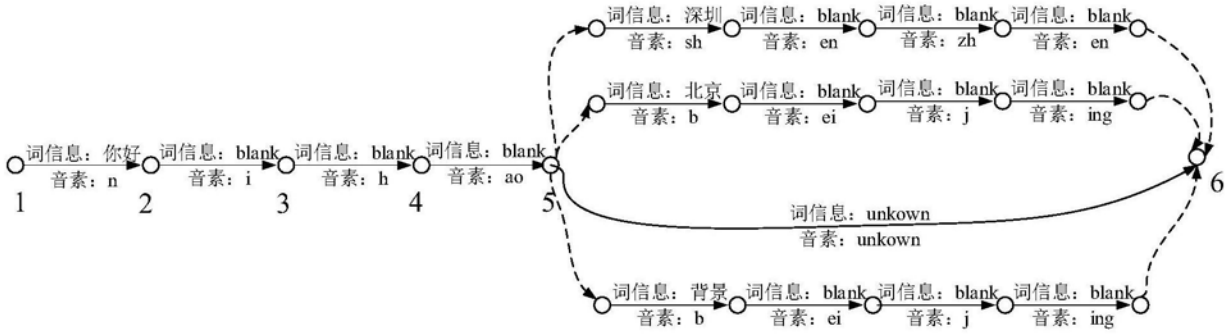

词信息：你好
音素：n
词信息：blank
音素：i
词信息：blank
音素：h
词信息：blank
音素：ao
1
2
3
4
5
6
词信息：深圳
音素：sh
词信息：blank
音素：en
词信息：blank
音素：zh
词信息：blank
音素：en
词信息：北京
音素：b
词信息：blank
音素：ei
词信息：blank
音素：j
词信息：blank
音素：ing
词信息：unkown
音素：unkown
词信息：背景
音素：b
词信息：blank
音素：ei
词信息：blank
音素：j
词信息：blank
音素：ing

图6

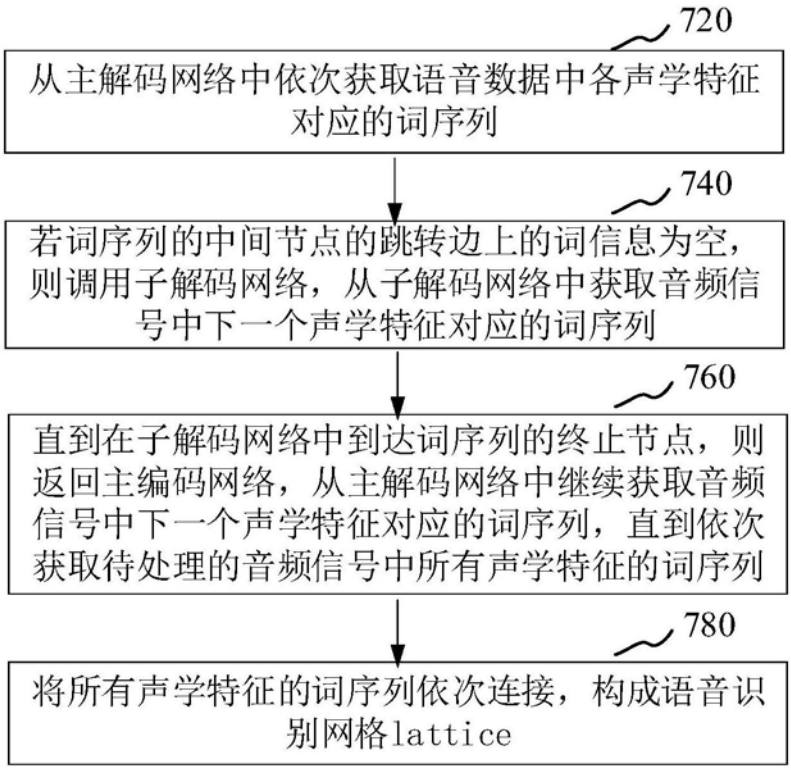

720
从主解码网络中依次获取语音数据中各声学特征对应的词序列
740
若词序列的中间节点的跳转边上的词信息为空，则调用子解码网络，从子解码网络中获取音频信号中下一个声学特征对应的词序列
760
直到在子解码网络中到达词序列的终止节点，则返回主编码网络，从主解码网络中继续获取音频信号中下一个声学特征对应的词序列，直到依次获取待处理的音频信号中所有声学特征的词序列
780
将所有声学特征的词序列依次连接，构成语音识别网格lattice

图7

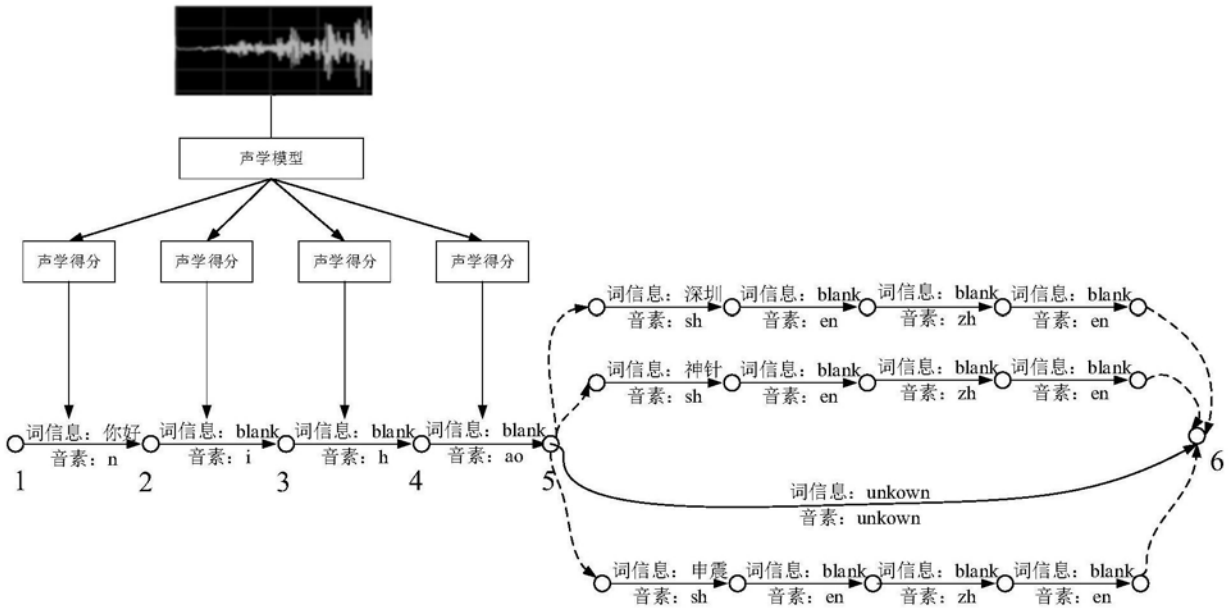

声学模型
声学得分
声学得分
声学得分
声学得分
词信息：你好
音素：n
词信息：blank
音素：i
词信息：blank
音素：h
词信息：blank
音素：ao
1
2
3
4
5
词信息：深圳
音素：sh
词信息：blank
音素：en
词信息：blank
音素：zh
词信息：blank
音素：en
词信息：神针
音素：sh
词信息：blank
音素：en
词信息：blank
音素：zh
词信息：blank
音素：en
词信息：unkown
音素：unkown
词信息：申震
音素：sh
词信息：blank
音素：en
词信息：blank
音素：zh
词信息：blank
音素：en
6

图8

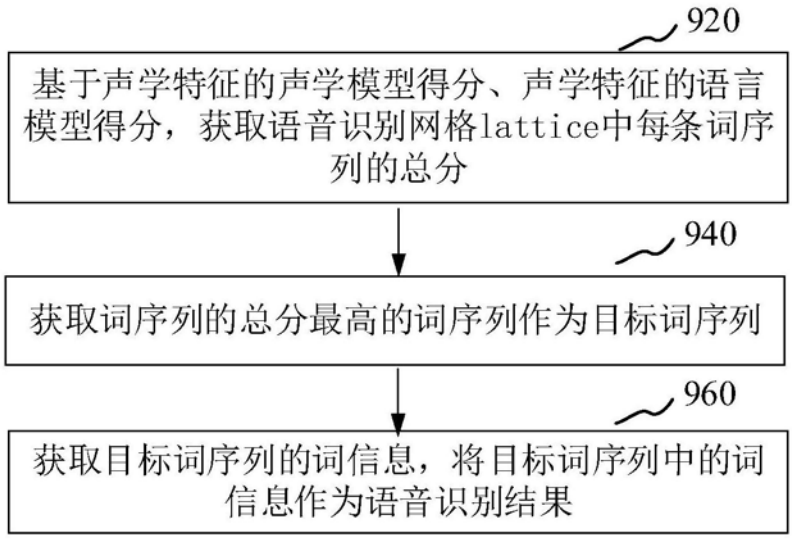

920
基于声学特征的声学模型得分、声学特征的语言模型得分，获取语音识别网格lattice中每条词序列的总分
940
获取词序列的总分最高的词序列作为目标词序列
960
获取目标词序列的词信息，将目标词序列中的词信息作为语音识别结果

图9

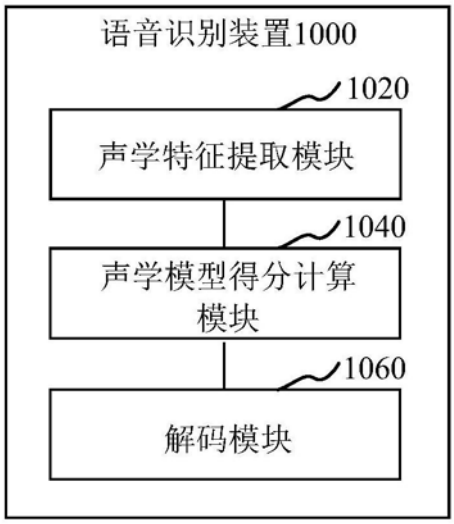

语音识别装置1000
1020
声学特征提取模块
1040
声学模型得分计算模块
1060
解码模块

图10

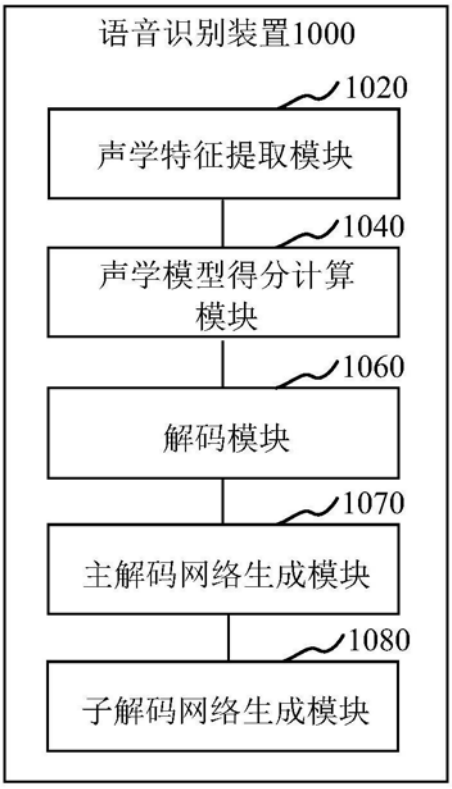

语音识别装置1000
1020
声学特征提取模块
1040
声学模型得分计算模块
1060
解码模块
1070
主解码网络生成模块
1080
子解码网络生成模块

图11

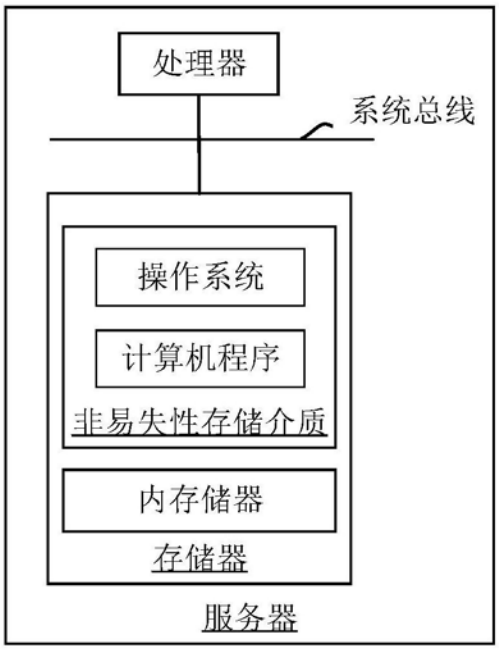


图12