(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109637537 B
(45)授权公告日 2020.06.30

(21)申请号 201811620403.0

(22)申请日 2018.12.28

(65)同一申请的已公布的文献号
申请公布号 CN 109637537 A

(43)申请公布日 2019.04.16

(73)专利权人 北京声智科技有限公司
地址 100086 北京市海淀区北四环西路67号3层306室

(72)发明人 杨程远 陈孝良 冯大航 苏少炜 常乐

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 任岩

(51)Int.Cl.
G10L 15/22(2006.01)
G10L 15/06(2013.01)
G10L 15/26(2006.01)

(56)对比文件
CN 107578769 A,2018.01.12,说明书第[0003],[0008],[0021]-[0024],[0068],[0075]-[0078],[0097]段,附图1-3.
CN 107247706 A,2017.10.13,说明书第[0089]段.
US 8543398 B1,2013.09.24,全文.
Ryohei Nakatani,Tetsuya Takiguchi,Yasuo Ariki.Two-step Correction of Speech Recognition Errors Based on N-gram and Long Contextual Information.《INTERSPEECH 2013》.2013,3747-3750.

审查员 李梦璐

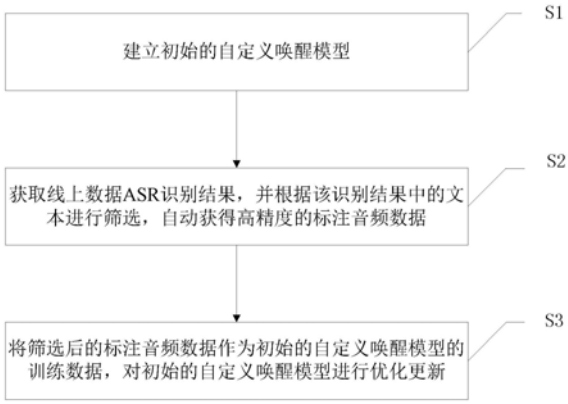
权利要求书1页 说明书6页 附图1页

(54)发明名称

一种自动获取标注数据优化自定义唤醒模型的方法

(57)摘要

本公开提供了一种自动获取标注数据优化自定义唤醒模型的方法包括:建立初始的自定义唤醒模型;获取线上数据ASR的识别结果,并根据该识别结果中的文本进行筛选,自动获得标注音频数据;将筛选后的标注音频数据作为初始的自定义唤醒模型的训练数据,对初始的自定义唤醒模型进行优化更新。基于N-Gram模型对ASR识别结果的文本成句概率进行筛选,从而自动获取标注音频数据,并将其作为自定义唤醒模型的训练数据,能够方便地实现自定义唤醒模型的优化训练。



CN 109637537 B

1.一种自动获取标注数据优化自定义唤醒模型的方法包括：

建立初始的自定义唤醒模型；

获取线上数据ASR的识别结果，并根据该识别结果中的文本进行筛选，自动获得标注音频数据；

将筛选后的标注音频数据作为初始的自定义唤醒模型的训练数据，对初始的自定义唤醒模型进行优化更新；

所述根据该识别结果中的文本进行筛选，自动获得标注音频数据，包括以下子步骤：

建立一个基础的N-Gram模型；

根据获取的线上数据ASR的识别结果，对线上数据ASR识别结果的文本进行分词；

分词后在所述N-Gram模型中依次查找文本中每一个N元词组，并计算文本中每一个N元词组出现的概率，计算完所有的N元词组的概率之后求均值，获得所述文本的成句概率；以及

筛选成句概率大于预定阈值的文本对应的标注音频，从而自动获得标注音频数据。

2.根据权利要求1所述的方法，所述建立初始的自定义唤醒模型包括：

采用已标注的音频输入到自定义唤醒模型进行训练，所述自定义唤醒模型输出用于确定是否进行语音唤醒的结果。

3.根据权利要求2所述的方法，其中，所述自定义唤醒模型由唤醒声学模型及解码器组成，自定义唤醒模型的唤醒词为用户自定义的。

4.根据权利要求3所述的方法，其中，自定义唤醒仅搜索唤醒声学模型的最高分。

5.根据权利要求4所述的方法，其中，自定义唤醒模型搜索唤醒声学模型中各种音素排序，每种音素排序对应一个分数。

6.根据权利要求1所述的方法，所述分词后在基础N-gram模型中依次查找文本中每一个N元词组时，如果找不到N元词组，则回退查找N-1元词组，若获取的为N-1元词组的成句概率，则对该文本的成句概率乘以折扣系数。

7.根据权利要求1所述的方法，所述筛选成句概率大于预定阈值的文本对应的标注音频包括：采用已经标注的预定数量的样本集合做测试，获取一个使得筛选后的数据的字准确率能够满足使用要求的成句概率的阈值；采用该阈值对文本进行筛选，获取筛选出的文本语句对应的音频，生成标注音频数据。

8.根据权利要求1所述的方法，其中，所述将筛选后的标注音频数据作为初始的自定义唤醒模型的训练数据包括：

在人工标注的音频数据的基础上，增加筛选后自动获取的标注音频数据，作为自定义唤醒模型训练数据。

一种自动获取标注数据优化自定义唤醒模型的方法

技术领域

[0001] 本公开涉及自动语音识别(Automatic Speech Recognition,简称ASR) 领域,尤其涉及一种基于N-Gram自动获取高精度标注数据优化自定义唤醒模型的方法。

背景技术

[0002] 随着信息和通信技术的发展,智能设备已经在日常生活中被广泛应用。智能音箱等智能设备可使用通过麦克风采集到的音频信号来提供服务,例如智能语音设备作为家庭场景中有效的人工智能交互入口。

[0003] 智能语音设备基于自动语音识别系统,自动语音识别系统由声学模型、语言模型、解码器三大部分构成。其中声学模型是由大量已经标注文本的音频的提特征之后通过DNN等方法训练得到的,大量领域相关、标注准确的音频是声学模型优化的主要途径。

[0004] 对于智能音箱领域,智能音箱获取到用户输入的音频数据后,需要先检测获取到的声音信号中是否包括唤醒词,如果包括唤醒词,便会激活语音识别系统,来对获取到的声音信号进行识别,从而根据所识别出的声音信号执行相应的操作,如果不包括唤醒词,则不激活语音识别系统,也就不会对获取到的声音信号进行识别。即语音唤醒技术是一种具有开关入口属性的功能,用户通过唤醒词的唤醒,可以发起人机交互的操作,即智能音箱只有被用户所说的唤醒词唤醒后,才会对用户接下来的声音信号进行识别。因此,对于智能音箱等智能设备,为了方便用户自定义唤醒词,需要自定义唤醒模型。

[0005] 与声学模型类似的,自定义唤醒模型同样需要大量的标注音频提高自定义唤醒词的覆盖率及准确性。在最初建立自定义唤醒模型之后,为了进一步提高唤醒词的覆盖率及准确性,还需要对自定义唤醒模型进行优化。

发明内容

[0006] (一)要解决的技术问题

[0007] 本公开提供了一种基于N-Gram自动获取高精度标注数据优化自定义唤醒模型的方法,以至少部分解决以上所提出的技术问题。

[0008] (二)技术方案

[0009] 根据本公开的一个方面,提供了一种自动获取标注数据优化自定义唤醒模型的方法包括:

[0010] 建立初始的自定义唤醒模型;

[0011] 获取线上数据ASR的识别结果,并根据该识别结果中的文本进行筛选,自动获得标注音频数据;

[0012] 将筛选后的标注音频数据作为初始的自定义唤醒模型的训练数据,对初始的自定义唤醒模型进行优化更新。

[0013] 在一些实施例中,所述建立初始的自定义唤醒模型包括:

[0014] 采用已标注的音频输入到自定义唤醒模型进行训练,所述自定义唤醒模型输出用

于确定是否进行语音唤醒的结果。

[0015] 在一些实施例中,所述自定义唤醒模型由唤醒声学模型及解码器组成,自定义唤醒模型的唤醒词为用户自定义的。

[0016] 在一些实施例中,自定义唤醒仅搜索唤醒声学模型的最高分。

[0017] 在一些实施例中,自定义唤醒模型搜索唤醒声学模型中各种音素排序,每种音素排序对应一个分数。

[0018] 在一些实施例中,所述根据该识别结果中的文本进行筛选,自动获得标注音频数据,包括以下子步骤:

[0019] 建立一个基础的N-Gram模型;

[0020] 根据获取的线上数据ASR的识别结果,基于所述基础的N-Gram模型对线上数据ASR识别结果的文本的成句概率进行计算;

[0021] 筛选成句概率大于预定阈值的文本对应的标注音频,从而自动获得标注数据。

[0022] 在一些实施例中,所述基于所述基础的N-Gram模型对线上数据ASR 识别结果的文本的成句概率进行计算包括:

[0023] 对线上数据ASR识别结果的文本进行分词;

[0024] 分词后在所述N-Gram模型中依次查找文本中每一个N元词组,并计算文本中每一个N元词组出现的概率,计算完所有的N元词组的概率之后求均值,获得所述文本的成句概率。

[0025] 在一些实施例中,所述分词后在基础N-gram模型中依次查找文本中每一个N元词组时,如果找不到N元词组,则回退查找N-1元词组,若获取的为N-1元词组的成句概率,则对该文本的成句概率乘以折扣系数。

[0026] 在一些实施例中,所述筛选成句概率大于预定阈值的文本对应的标注音频包括:采用已经标注的预定数量的样本集合做测试,获取一个使得筛选后的数据的字准确率能够满足使用要求的成句概率的阈值;获取筛选出的文本语句对应的音频,生成标注音频数据。

[0027] 在一些实施例中,所述将筛选后的标注音频数据作为初始的自定义唤醒模型的训练数据包括:

[0028] 在人工标注音频数据的基础上,增加筛选后自动获取的标注音频数据,作为自定义唤醒模型训练数据。

[0029] (三)有益效果

[0030] 从上述技术方案可以看出,本公开基于N-Gram自动获取高精度标注数据优化自定义唤醒模型的方法至少具有以下有益效果:

[0031] 基于N-Gram模型对ASR识别结果的文本成句概率进行筛选,从而自动获取高精度的标注音频数据,并将其作为自定义唤醒模型的训练数据,能够方便地实现自定义唤醒模型的优化训练。

附图说明

[0032] 图1为本公开实施例基于N-Gram自动获取高精度标注数据优化自定义唤醒模型的方法的流程图。

具体实施方式

[0033] 本公开提供了一种基于N-Gram自动获取高精度标注数据优化自定义唤醒模型的方法。为使本公开的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本公开进一步详细说明。

[0034] 本公开某些实施例于后方将参照所附附图做更全面性地描述，其中一些但并非全部的实施例将被示出。实际上，本公开的各种实施例可以由许多不同形式实现，而不应被解释为限于此处所阐述的实施例；相对地，提供这些实施例使得本公开满足适用的法律要求。

[0035] 在本公开的一个示例性实施例中，提供了一种基于N-Gram自动获取高精度标注数据优化自定义唤醒模型的方法。图1为本公开实施例基于 N-Gram自动获取高精度标注数据优化自定义唤醒模型的方法的流程图。如图1所示，本公开基于N-Gram自动获取高精度标注数据优化自定义唤醒模型的方法包括：

[0036] S1，建立初始的自定义唤醒模型；

[0037] S2，获取线上数据ASR识别结果，并根据该识别结果中的文本进行筛选，自动获得高精度的标注音频数据；

[0038] S3，将筛选后的标注音频数据作为初始的自定义唤醒模型的训练数据，对初始的自定义唤醒模型进行优化更新。

[0039] 所述步骤S1中，首先采用大量已标注的音频输入到自定义唤醒模型进行训练，所述已标注的音频一般为通用的语音训练数据。所述自定义唤醒模型输出用于确定是否进行语音唤醒的结果。其中，所述自定义唤醒模型即唤醒词为用户自定义的，而非指定的唤醒词。

[0040] 通常，初始唤醒模型的性能并不能达到最优水平，在实际应用过程中，由于初始唤醒模型性能不高，导致唤醒成功的语音数据中可能存在误唤醒数据，例如，背景噪音、人声干扰、与唤醒词相近发音的非唤醒词等，均可能误唤醒智能终端。因此，需要在使用过程中不断进行模型优化，来提高模型的识别准确率。

[0041] 而智能音箱的线上用户数据数以百万、千万计，线上数据是非常有价值的，其中，线上数据主要包括：用户真实的跟智能音箱对话的声音(音频)、ASR系统识别出来的文本内容(音频对应的识别文本)。真实的用户数据更具有相关性，能够获取大量的准确性高、相关性高的文本内容。但线上数据识别结果存在一定差错率，为了满足准确性要求，需要对线上数据进行筛选。

[0042] 所述步骤S2中对数据进行筛选，包括以下子步骤：

[0043] S201，首先建立一个基础的N-Gram模型；

[0044] S202，获取线上数据ASR识别结果，基于所述基础的N-Gram模型对线上数据ASR识别结果的文本的成句概率进行计算；

[0045] S203，筛选成句概率大于预定阈值的文本对应的标注音频，从而自动获得高精度的标注数据。

[0046] 具体地，所述N-Gram模型即N元文法模型。所述N-Gram模型是大词汇连续语音识别中常用的一种统计语言模型，利用上下文中相邻词间的搭配信息，统计词频，计算出具有最大概率的句子，或者进行文本拼写检查。在拼写检查的应用中，N-Gram模型由于存在数据的稀疏性，需要加上平滑算法才能表现出良好的效果。平滑技术(smoothing)通过提高低概率

或零概率,降低高概率,使统计语言模型可求,从而解决零概率问题,一般地,可以采用回退(backoff)和插值(interpolate)两种方法来实现平滑。

[0047] 最初建立基础的N-Gram模型时可以以大量语料进行模型训练,提高覆盖度和普遍性。通过开源工具包Srilm和IRSTLM以C++作为开发语言,建立N-Gram模型速度表现也较好。

[0048] 采用神经网络模型进行分词,通过使用神经网络训练而得的分词词典,以前后双向匹配取最少量分词结果的方式进行分词。例如:我是中国人;前向分词结果:我|是|中国|人;后向结果是:我是|中国人;那么会选择后向结果。

[0049] 在分词后,在基础的N-Gram模型中依次查找文本中每一个N元词组出现的概率。具体地,所述N-Gram模型基于马尔科夫假设:第N个词的出现只与前面N-1个词相关,而与其它任何词都不相关,整句的概率就是各个词出现概率的乘积。再通过所有的N元词组的概率求均值,获得所述文本的成句概率。优选地,基于N元语法的统计语言模型最常采用三元语法或二元语法。

[0050] 在获得文本的成句概率后,根据设定的阈值,对线上数据ASR识别结果的文本进行筛选,即可完成文本的筛选。其中,所述阈值使得筛选后生成的高精度标注数据的字准确率能够满足使用要求。

[0051] 所述步骤S3中,将筛选后的文本对应的标注音频数据作为初始的自定义唤醒模型的训练数据,对初始的自定义唤醒模型进行优化更新。所述标注音频数据包括文本及对应的音频。

[0052] 以下结合具体实施例对本公开用基于N-Gram自动获取高精度标注数据优化自定义唤醒模型的方法的各个步骤进行详细说明。

[0053] 所述步骤S1中,首先建立初始的自定义唤醒模型,采用大量已标注的音频作为输入进行训练,所述初始的自定义唤醒模型输出用于确定是否进行语音唤醒的结果。一般地,用于训练初始的自定义唤醒模型的语音数据可以来自数据库中或网络上通用的已标注音频数据。

[0054] 自定义唤醒词可以由用户输入,而且用户自定义的唤醒词可以为一个也可以为多个,比如说,预定的唤醒词可以为简单的词组、数字、字母,或其任意组合,用户可以根据自己的实际需求或个人喜好进行设置,比如说唤醒词可以为“打电话”、“打开灯”、“关上灯”等词。自定义唤醒过程不需要将自定义的唤醒词识别成字;自定义唤醒模型只到音素级别,可以理解成只识别到拼音级别。一般的,ASR包括声学模型、语言模型、解码器;声学模型的输入是声学特征,输出是音素;语言模型将音 素序列转换成文字序列;解码器就是一个搜索排序的功能,从声学模型和语言模型里搜索各种排序,每个排序都有一个分数,解码器将声学模型和语言模型的排序分数相加,选最高分作为asr识别结果;而自定义唤醒只包括唤醒声学模型及解码器,自定义唤醒只搜索唤醒声学模型的最高分。

[0055] 在训练过程中,将唤醒成功的语音数据视为正例语音数据,将唤醒失败的语音数据视为反例语音数据,基于区分性准则对当前唤醒模型进行训练优化。

[0056] 所述步骤S2中,首先要建立一个基础的N-Gram模型,N-Gram模型可以计算一串词语的成句概率,举例来说就是:“我爱美丽的花朵”肯定要比“我爱光明的花朵”更像是一句合理的话,所谓的合理就是“我爱美丽的花朵”这三个词的成句概率更高;或者说“随波逐流”要比“逐流随波”更加常见,所谓常见就是在一个大量文本的数据集(比如1千万个某论

坛的网页数据)中“随波逐流”要比“逐流随波”出现的次数多。

[0057] 具体地,所述N-Gram模型是基于统计的模型,其中有大量文本分词后的词频的概率统计,以及词语和词语同时出现的概率统计,还有对数据稀疏问题做平滑的回退概率,可以表征文本的上下文关系。本实施例中,使用三元文法模型,就是最多统计一个词前面两个词的概率,即某三个词出现的概率Pi(i=1,2,3…)。

[0058] 在拿到一段音频的识别文本后,比如是“给我播放一首刘若英的后来”,会首先将这段话分词,比如分完词之后是:<s>给我|播放|一首|刘若英|的| 后来</s>;其中<s>和</s>分别是一段话的开头和结尾标识。分完词之后会在基础三元模型中依次查找文本中每一个三元词组,在此可以认为是对分词的结果进行了有顺序的排列组合,例如上述文本中存在的三元词组包括:给我|播放|一首、给我|播放|刘若英、给我|播放1后来、播放|一首|刘若英|、刘若英|的|后来等等。

[0059] 根据基础N-Gram模型判断文本成句概率。本实施例中采用基础三元模型,查找分词之后在基础三元模型中每一个三元词组出现的概率Pi (i=1,2,3…);在此,需要要求基础N-Gram模型尽可能要全面丰富、内容相关,所述基础N-gram模型中包括已经计算好的所有已经出现过的词组的共现概率,以及数据稀疏情况下的回退概率。如果找不到三元就回退到二元,获取二元组的出现概率Pi’(i=1,2,3…),同时需要对该文本的成句概率乘以折扣系数Q,即成句概率为Pi’*Q。优选地,采用的折扣系数为 0.7～0.8,在该范围里对文本成句概率结果影响不大。

[0060] 计算完文本中所有的三元词组的概率之后求均值,就是该文本语句的成句概率;对于具有n个三元词组的文本,所述成句概率可以表示为:

[0061] P=(P1+P2+...+Pn)/n;

[0062] 其中,P1为文本语句中第一个三元词组的成句概率;P2为文本语句中第二个三元词组的成句概率,……,Pn为文本语句中第n个三元词组的成句概率。

[0063] 获取文本语句的成句概率后,使用确定成句概率的阈值对线上数据进行筛选。所述阈值可以预先确定,本实施例中,通过使用4万条已经标注的小样本集合做测试,找到一个成句概率的阈值,使得筛选后的数据的字准确率能够满足使用要求。通过使用该阈值对数以百万计的线上数据做筛选,获取成句概率高于阈值的文本语句。优选地,筛选之前会把高频句子进行删减,并会去除单字。

[0064] 进一步地,获取筛选出的文本语句对应的音频,生成高精度标注音频数据。

[0065] 本实施例所述步骤S3中,将筛选后的相关标注音频用于自定义唤醒模型训练,优选地,可以人工标注数据的基础上,增加自动标注数据,作为训练数据。在一具体实施例中,所述步骤S3是在1600小时人工标注的音频数据基础上,增量了筛选后的1500小时数据;自定义唤醒测试识别率提高1.8个点,使用筛选数据的弥补了数据量的不足。

[0066] 至此,已经结合附图对本公开实施例进行了详细描述。需要说明的是,在附图或说明书正文中,未绘示或描述的实现方式,均为所属技术领域中普通技术人员所知的形式,并未进行详细说明。此外,上述对各元件和方法的定义并不仅限于实施例中提到的各种具体结构、形状或方式,本领域普通技术人员可对其进行简单地更改或替换。

[0067] 此外,除非特别描述或必须依序发生的步骤,上述步骤的顺序并无限制于以上所列,且可根据所需设计而变化或重新安排。并且上述实施例可基于设计及可靠度的考虑,彼

此混合搭配使用或与其他实施例混合搭配使用,即不同实施例中的技术特征可以自由组合形成更多的实施例。

[0068] 在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备固有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描述,构造这类系统所要求的结构是显而易见的。此外,本公开也不针对任何特定编程语言。应当明白,可以利用各种编程语言实现在此描述的本公开的内容,并且上面对特定语言所做的描述是为了披露本公开的最佳实施方式。

[0069] 本公开可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。本公开的各个部件实施例可以以硬件实现,或者以在一个或者多个处理器上运行的软件模块实现,或者以它们的组合实现。本领域的技术人员应当理解,可以在实践中使用微处理器或者数字信号处理器 (DSP)来实现根据本公开实施例的相关设备中的一些或者全部部件的一些或者全部功能。本公开还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序(例如,计算机程序和计算机程序产品)。这样的实现本公开的程序可以存储在计算机可读介质上,或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到,或者在载体信号上提供,或者以任何其他形式提供。

[0070] 本领域那些技术人员可以理解,可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件,以及此外可以把它们分成多个子模块或子单元或子组件。除了这样的特征和/ 或过程或者单元中的至少一些是相互排斥之外,可以采用任何组合对本说明书(包括伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述,本说明书(包括伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。并且,在列举了若干装置的单元权利要求中,这些装置中的若干个可以是通过同一个硬件项来具体体现。

[0071] 类似地,应当理解,为了精简本公开并帮助理解各个公开方面中的一个或多个,在上面对本公开的示例性实施例的描述中,本公开的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而,并不应将该公开的方法解释成反映如下意图:即所要求保护的本公开要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说,如下面的权利要求书所反映的那样,公开方面在于少于前面公开的单个实施例的所有特征。因此,遵循具体实施方式的权利要求书由此明确地并入该具体实施方式,其中每个权利要求本身都作为本公开的单独实施例。

[0072] 以上所述的具体实施例,对本公开的目的、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,以上所述仅为本公开的具体实施例而已,并不用于限制本公开,凡在本公开的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本公开的保护范围之内。

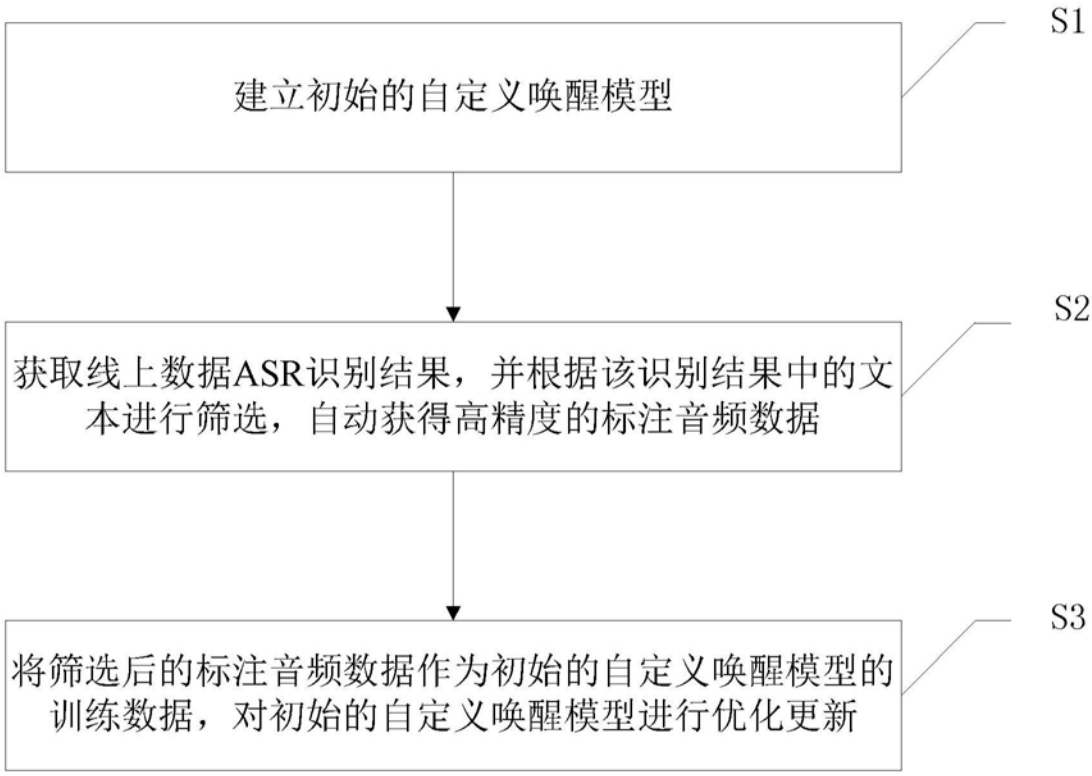

S1
建立初始的自定义唤醒模型
S2
获取线上数据ASR识别结果，并根据该识别结果中的文本进行筛选，自动获得高精度的标注音频数据
S3
将筛选后的标注音频数据作为初始的自定义唤醒模型的训练数据，对初始的自定义唤醒模型进行优化更新

图1