(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 104464723 B
(45)授权公告日 2018.03.20

(21)申请号 201410782284.4
(22)申请日 2014.12.16
(65)同一申请的已公布的文献号
申请公布号 CN 104464723 A
(43)申请公布日 2015.03.25
(73)专利权人 科大讯飞股份有限公司
地址 230088 安徽省合肥市高新开发区望江西路666号
(72)发明人 张凯 陈盛
(74)专利代理机构 北京维澳专利代理有限公司 11252
代理人 王立民 姜溯洲
(51)Int.Cl.
G10L 15/04(2013.01)
G10L 15/26(2006.01)

审查员 李召卿

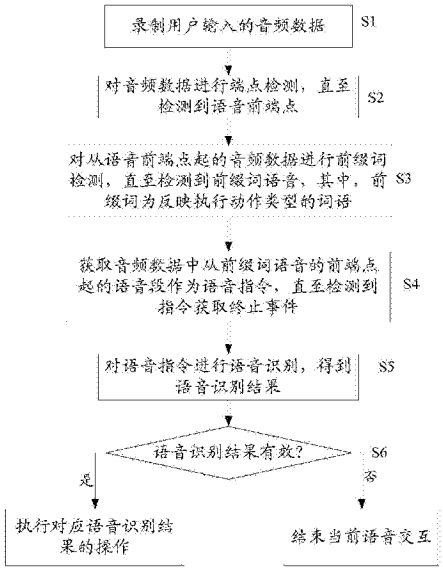
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种语音交互方法及系统

(57)摘要

本发明公开了一种语音交互方法及系统，该方法包括录制用户输入的音频数据；对音频数据进行端点检测，直至检测到语音前端点；对从语音前端点起的音频数据进行前缀词检测，直至检测到前缀词语音，该前缀词为反映执行动作类型的词语；获取音频数据中从前缀词语音的前端点起的语音段作为语音指令；对语音指令进行语音识别；如果语音识别结果有效则执行对应语音识别结果的操作。本发明的方法及系统由于将音频数据中从前缀词语音的前端点起的语音段作为语音指令，并将反映执行动作类型的词语作为前缀词，因此实现了前缀词与语音指令间的有机结合，可以有效避免出现因强制切分语音指令带来的无法获得有效语音识别结果的问题，提高了语音交互的效率。



CN 104464723 B

1.一种语音交互方法,其特征在于,包括:

录制用户输入的音频数据;

对所述音频数据进行端点检测,直至检测到语音前端点;

对从所述语音前端点起的音频数据进行前缀词检测,直至检测到前缀词语音,其中,所述前缀词为反映执行动作类型的词语,并且所述前缀词与用于表明用户意图的语音指令结合在一起;

获取所述音频数据中从所述前缀词语音的前端点起的语音段作为语音指令,直至检测到指令获取终止事件;

对所述语音指令进行语音识别,得到语音识别结果;

判断所述语音识别结果是否有效,如果有效则执行对应所述语音识别结果的操作。

2.根据权利要求1所述的方法,其特征在于,所述方法还包括:

在对所述音频数据进行端点检测之前,对所述音频数据进行降噪处理。

3.根据权利要求1所述的方法,其特征在于,所述对从所述语音前端点起的音频数据进行前缀词检测包括:

基于包括前缀词模型和垃圾模型的并行搜索网络,检测从所述语音前端点起的音频数据中是否存在所述前缀词语音。

4.根据权利要求1所述的方法,其特征在于,所述判断所述语音识别结果是否有效包括:

判断命令词网络中是否存在与所述语音识别结果相匹配的命令词,如存在,则判定所述语音识别结果有效。

5.根据权利要求1至4中任一项所述的语音交互方法,其特征在于,所述指令获取终止事件包括:所述语音段结束和所述语音段已持续设定时间。

6.一种语音交互系统,其特征在于,包括:

录音模块,用于录制用户输入的音频数据;

端点检测模块,用于对所述音频数据进行端点检测,直至检测到语音前端点;

前缀词检测模块,用于对从所述语音前端点起的音频数据进行前缀词检测,直至检测到前缀词语音,其中,所述前缀词为反映执行动作类型的词语,并且所述前缀词与用于表明用户意图的语音指令结合在一起;

语音活动检测模块,用于获取所述音频数据中从所述前缀词语音的前端点起的语音段作为语音指令,直至检测到指令获取终止事件;

语音识别模块,用于对所述语音指令进行语音识别,得到语音识别结果;

判断模块,用于判断所述语音识别结果是否有效;以及,

执行模块,用于执行有效的语音识别结果对应的操作。

7.根据权利要求6所述的系统,其特征在于,所述系统还包括:

降噪模块,分别与所述录音模块及所述端点检测模块连接,用于对所述录音模块录制的音频数据进行降噪处理,并将降噪处理后的音频数据传送给所述端点检测模块。

8.根据权利要求6所述的系统,其特征在于,所述前缀词检测模块具体用于基于包括前缀词模型和垃圾模型的并行搜索网络,检测从所述语音前端点起的音频数据中是否存在所述前缀词语音。

9.根据权利要求6所述的系统，其特征在于，所述判断模块具体用于判断命令词网络中是否存在与所述语音识别结果相匹配的命令词，如存在，则判定所述语音识别结果有效。

10.根据权利要求6至9中任一项所述的系统，其特征在于，所述指令获取终止事件包括：所述语音段结束和所述语音段已持续设定时间。

一种语音交互方法及系统

技术领域

[0001] 本发明涉及语音交互领域，尤其涉及一种语音交互方法及系统。

背景技术

[0002] 为了避免手机等移动设备在待机时将周边的说话噪音误识别为语音指令，用户在每次启动移动设备的语音交互功能时，移动设备均需要完成以下操作：1、录制用户输入的音频数据；2、获取音频数据进行唤醒检测，直至唤醒成功；3、于唤醒成功后提示用户输入语音指令；4、于提示用户输入语音指令后，再次录制用户输入的音频数据；5、获取再次录制的音频数据中的语音段作为语音指令；6、对语音指令进行语音识别，得到语音识别结果；7、确定语音识别结果是否有效，如果有效则执行语音识别结果。对应地，用户在每次启动移动设备的语音交互功能时，均需要完成以下操作：1、说出唤醒词，以唤醒移动设备；2、在移动设备提示用户输入语音指令时，说出语音指令，例如说出“打电话给张三”时。由此可见，该种语音交互方法具有使用便捷性较差的缺陷。

[0003] 为了解决上述语音交互方法存在的使用便捷性较差的问题，目前还提出了一种基于唤醒词的语音交互方法，该种语音交互方法是在唤醒成功后直接处理用户在说出唤醒词后连续说出的语音指令。与该种语音交互方法相对应，用户需要完成的操作是连续说出唤醒词和语音指令，例如，对于要“打电话给张三”的应用，用户需要说出“语点通，打电话给张三”，其中的“语点通”即为预先设定的固定唤醒词，而“打电话给张三”即为语音指令。该种语音交互方法虽然在使用便捷性上具有一定的优势，但是，用户通常都是连续说话，唤醒词与后面的语音指令会顺连在一起，因此，这种将音频数据中于唤醒成功起的语音段作为语音指令的强制切分方式，很可能导致语音指令不完整，进而导致语音识别模块无法获得有效的语音识别结果，降低了语音识别模块的识别准确率，这就在一定程度上降低了语音交互的效率。另外，该种语音交互方法仅针对固定的唤醒词起作用，用户需要硬性记忆设定的唤醒词，否则将无法开始整个语音交互过程，因此，该种语音交互方法的使用便捷性仍有待进一步提高。

发明内容

[0004] 本发明实施例的目的在于克服现有语音交互方法存在的语音交互效率较低的问题，提供了一种高效的基于前缀词的语音交互方法。

[0005] 为实现上述目的，本发明采用的技术方案为：一种语音交互方法，包括：

[0006] 录制用户输入的音频数据；

[0007] 对所述音频数据进行端点检测，直至检测到语音前端点；

[0008] 对从所述语音前端点起的音频数据进行前缀词检测，直至检测到前缀词语音，其中，所述前缀词为反映执行动作类型的词语，并且所述前缀词与用于表明用户意图的语音指令结合在一起；

[0009] 获取所述音频数据中从所述前缀词语音的前端点起的语音段作为语音指令，直至

检测到指令获取终止事件；

[0010] 对所述语音指令进行语音识别，得到语音识别结果；

[0011] 判断所述语音识别结果是否有效，如果有效则执行对应所述语音识别结果的操作。

[0012] 优选的是，所述方法还包括：

[0013] 在对所述音频数据进行端点检测之前，对所述音频数据进行降噪处理。

[0014] 优选的是，所述对从所述语音前端点起的音频数据进行前缀词检测包括：

[0015] 基于包括前缀词模型和垃圾模型的并行搜索网络，检测从所述语音前端点起的音频数据中是否存在所述前缀词语音。

[0016] 优选的是，所述判断所述语音识别结果是否有效包括：

[0017] 判断命令词网络中是否存在与所述语音识别结果相匹配的命令词，如存在，则判定所述语音识别结果有效。

[0018] 优选的是，所述指令获取终止事件包括：所述语音段结束和所述语音段已持续设定时间。

[0019] 为了实现上述目的，本发明采用的技术方案为：一种语音交互系统，包括：

[0020] 录音模块，用于录制用户输入的音频数据；

[0021] 端点检测模块，用于对所述音频数据进行端点检测，直至检测到语音前端点；

[0022] 前缀词检测模块，用于对从所述语音前端点起的音频数据进行前缀词检测，直至检测到前缀词语音，其中，所述前缀词为反映执行动作类型的词语，并且所述前缀词与用于表明用户意图的语音指令结合在一起；

[0023] 语音活动检测模块，用于获取所述音频数据中从所述前缀词语音的前端点起的语音段作为语音指令，直至检测到指令获取终止事件；

[0024] 语音识别模块，用于对所述语音指令进行语音识别，得到语音识别结果；

[0025] 判断模块，用于判断所述语音识别结果是否有效；以及，

[0026] 执行模块，用于执行有效的语音识别结果对应的操作。

[0027] 优选的是，所述系统还包括：

[0028] 降噪模块，分别与所述录音模块及所述端点检测模块连接，用于对所述录音模块录制的音频数据进行降噪处理，并将降噪处理后的音频数据传送给所述端点检测模块。

[0029] 优选的是，所述前缀词检测模块具体用于基于包括前缀词模型和垃圾模型的并行搜索网络，检测从所述语音前端点起的音频数据中是否存在所述前缀词语音。

[0030] 优选的是，所述判断模块具体用于判断命令词网络中是否存在与所述语音识别结果相匹配的命令词，如存在，则判定所述语音识别结果有效。

[0031] 优选的是，所述指令获取终止事件包括：所述语音段结束和所述语音段已持续设定时间。

[0032] 本发明的有益效果在于，本发明的语音交互方法及系统由于将音频数据中从前缀词语音的前端点起的语音段作为语音指令，并将例如是“打电话给”、“发短信给”、“打开QQ”等反映执行动作类型的词语作为前缀词，因此实现了前缀词与语音指令间的有机结合，这不仅可以有效避免出现因强制切分语音指令带来的无法获得有效语音识别结果的问题，提高了语音交互的效率，而且这种将符合常规语言习惯的词语作为前缀词的方式，使用户无

需硬性记忆固定的唤醒词，只需按照常规语言习惯说出需要执行的动作即可实现语音交互的唤醒和动作的执行，进而进一步提高了语音交互的使用便捷性。

附图说明

[0033] 图1示出了根据本发明所述语音交互方法的一种实施方式的流程图；

[0034] 图2示出了根据本发明所述语音交互系统的一种实施结构的方框原理图。

具体实施方式

[0035] 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能解释为对本发明的限制。

[0036] 本发明为了解决现有语音交互方法存在的因对语音指令进行强制切分而影响语音交互效率的问题，提供一种更为高效的语音交互方法，如图1所示，该方法包括如下步骤：

[0037] 步骤S1：录制用户输入的音频数据。

[0038] 在此，可将录制的音频数据存储在固定长度的循环缓冲区中，并记录存储地址，以供后续步骤获取该音频数据。

[0039] 步骤S2：对音频数据进行端点检测，直至检测到语音前端点。

[0040] 该语音前端点就是非语音段到语音段的边界帧，进行音频数据处理时，首先对音频数据进行分帧，然后对音频数据的每帧数据计算能量特征，能量特征超过设定数值就认为该帧数据是语音，否则是非语音。

[0041] 在此，音频数据会随着录音的进行不断被存储在循环缓冲区中，而随着音频数据的不断存储，便可不断从循环缓冲区中获取音频数据进行端点检测，因此，该对音频数据进行端点检测的动作基本可与将录制的音频数据存储在循环缓冲区中的动作同步进行，以提高处理效率。

[0042] 步骤S3：对从语音前端点起的音频数据进行前缀词检测，直至检测到前缀词语音，其中，所述前缀词为反映执行动作类型的词语，以能够将用于唤醒语音交互的前缀词与用于表明用户意图的语音指令有机地结合在一起。该反映执行动作类型的词语例如是“打电话给”、“发短信给”、“打开QQ”、“打开微信”等符合常规语言习惯的词语。

[0043] 该前缀词检测的主要作用为判断是否唤醒语音交互操作，如果检测到前缀词语音，则启动语音识别，以按照用户意图执行相应动作。

[0044] 该前缀词检测的方法例如可包括如下步骤：

[0045] 步骤S31，声学特征提取：提取音频信息(通常以语音段为单位进行前缀词检测)中具有区分性的、并且是基于人耳听觉特性提取的特征，通常选取语音识别中用到的MFCC(Mel-Frequency Cepstrum Coefficient，美尔频率倒谱系数)特征作为声学特征。

[0046] 步骤S32，前缀词检测：将提取得到的声学特征，采用训练的声学模型在前缀词检测网络上计算声学得分，如果声学得分最优的路径中包含要检测的前缀词，则确定已检出前缀词，否则回到步骤S31继续提取声学特征。

[0047] 在上述步骤S31和步骤S32的基础上，为了降低前缀词的误检率，还可以在确定已检出前缀词后执行以下步骤S33。

[0048] 步骤S33，前缀词确认：将提取得到的声学特征，采用训练的声学模型在前缀词确认网络上进行前缀词确认，得到最终确认得分；判断该检出的前缀词是否为真实的前缀词，即将该前缀词的最终确认得分和预先设定的门限进行比较，如果最终确认得分大于等于门限，则认为该前缀词是真实的前缀词，语音唤醒成功；如果最终确认得分小于门限，则认为该前缀词为虚假的前缀词，重新回到步骤S31继续提取声学特征。

[0049] 在此，可将符合常规语言习惯的反映执行动作类型的词语增加在前缀词检测网络和前缀词确认网络中，另外，本发明的方法还支持用户根据个人语言习惯，将反映执行动作类型的词语增加在前缀词检测网络和前缀词确认网络的操作。这使得本发明的方法不再受限于固定唤醒词，进一步提高了本发明的应用便捷性。

[0050] 上述前缀词检测网络的实现方法可采用最优得分路径计算得出，最优得分路径的计算公式是：

[0051] 现用X代表从音频数据中提取的声学特征向量，W代表得分最大的最优词序列；条件概率P(X|W)为声学模型得分，通过训练好的声学模型计算得到；先验概率P(W)为语言模型得分，即为对不同的声学模型所加的PenaltyP(X)为全概率，当声学模型和前缀词检测网络确定下来后即是定值。在此基础上，前缀词确认网络的实现方法是：

[0052] a)将检出的前缀词解码到音素一级，并记录所有的得分：

[0053] $(Score_{phone1}, Score_{phone2}, \cdots, Score_{phoneN})$，其中N为前缀词中总的音素个数，$Score_{phone1}, Score_{phone2}, \cdots, Score_{phoneN}$分别表示该前缀词中各音素的解码得分。

[0054] b)计算得到前缀词每个音素的确认得分，计算方式如下：

[0055] $$CM_{phonei} = (Score_{phonei} - \sum_{k=K_{istart}}^{K_{iend}} Score_{framek}) / (K_{iend} - K_{istart})$$

[0056] 其中K_{istart}和K_{iend}分别为第i个音素的起始时间和结束时间；CM_{phonei}表示第i个音素的确认得分，下标phonei表示第i个音素，$Score_{phonei}$如上面所示第i个音素的解码得分，$Score_{framek}$表示使用前缀词确认网络解码得到的第k帧的得分。

[0057] c)计算得到该前缀词的最终确认得分CM_{word}，计算方式如下所示：

[0058] $$CM_{word} = \frac{1}{N}\sum_{i=1}^{N} CM_{phonei}$$。

[0059] 为了提高前缀词检测效率及准确度，上述声学模型的训练可分为两部分，分别为前缀词模型和垃圾模型(即filler模型)；前缀词模型可采用传统的语音识别中的声学模型训练方法，选取数据库，利用基于MLE(Maximum Likelihood Estimation，最大似然估计)和MPE(Minimum Phone Error，最小音素错误)区分性训练准则下得到；而垃圾模型则用于吸收除前缀词之外的无关语音。因此，上述步骤S3中对从语音前端点起的音频数据进行前缀词检测可进一步包括：基于包括前缀词模型和垃圾模型的并行搜索网络，检测从语音前端点起的音频数据中是否存在前缀词语音。

[0060] 本领域技术人员应该理解的是，本发明也可以采用语音交互领域中惯常采用的其他字词检测手段检测前缀词语音，对此本发明实施例不做限定。

[0061] 步骤S4：获取音频数据中从前缀词语音的前端点起的语音段作为语音指令，直至检测到指令获取终止事件，以实现前缀词与语音指令的有机结合。

[0062] 在此，步骤S1的操作在检测到前缀词语音(即唤醒成功)后无中断地继续进行，而获取语音指令的动作由唤醒成功触发，该步骤即是在唤醒成功后直接从循环缓冲区中获取音频数据中的语音段。

[0063] 为了便于获取该语音段，可在检测到前缀词语音后，记录前缀词语音的后端点在循环缓冲区中的储存地址及前缀词语音的长度，这样，即可计算得到前缀词语音的前端点在循环缓冲区中的储存地址，从而可以准确获取音频数据中从前缀词语音的前端点起的语音段。

[0064] 步骤S5：对语音指令进行语音识别，得到语音识别结果。

[0065] 步骤S6：判断语音识别结果是否有效，如果有效则执行对应语音识别结果的操作；如果无效则结束本次语音交互，在此，可以提醒用户交互失败，并提醒用户再次输入正确的语音指令。

[0066] 本发明的语音交互方法由于将音频数据中从前缀词语音的前端点起的语音段作为语音指令，并将反映执行动作类型的词语作为前缀词，因此实现了前缀词与语音指令间的有机结合，这不仅可以有效避免出现因强制切分语音指令带来的无法获得有效语音识别结果的问题，提高了语音交互的效率，而且这种将符合常规语言习惯的词语作为前缀词的方式，使用户无需硬性记忆固定的唤醒词，只需按照常规语言习惯说出需要执行的动作即可实现语音交互的唤醒和动作的执行，进而进一步提高了语音交互的使用便捷性。

[0067] 为了提高前端点检测、前缀词检测及语音识别的准确度，及提高本发明语音交互方法的抗干扰能力，本发明的方法还可以在对音频数据进行端点检测之前，对音频数据进行降噪处理，得到干净音频数据，对此，上述步骤S3具体是对从语音前端点起的干净音频数据进行前缀词检测，上述步骤S4具体是获取干净音频数据中从前缀词语音的前端点起的语音段作为语音指令。

[0068] 上述步骤S6中判断所述语音识别结果是否有效可进一步包括如下步骤：

[0069] 步骤S61：加载命令词网络。

[0070] 本发明的方法支持用户根据需要扩充命令词网络的操作。

[0071] 步骤S62：判断命令词网络中是否存在与语音识别结果相匹配的命令词，如存在，则判定所述语音识别结果有效。

[0072] 在此，可通过计算语音识别结果与各命令词之间的相似度得到语音识别结果与各命令词之间的匹配度得分，如果匹配度得分高于或者等于设定阈值，则认为语音识别结果有效，否则认为语音结果无效。

[0073] 上述指令获取终止事件可根据需要设定，例如包括：语音段结束和语音段已持续设定时间。因此，在检测到前缀词语音后，可同时对音频数据中从前缀词语音的前端点起的语音段进行语音识别、后端点检测及持续时间计时。本领域技术人员可以根据实际应用场合将该设定时间设置为固定值，或者将该设定时间设置为可由用户输入确定，通常情况下，该设定时间在800ms至2000ms的范围内选择，例如选择为1000ms。上述语音段结束表示检测到语音段的后端点。如果在语音段持续设定时间时还未检测到后端点，则同样认为语音段结束。在此，每个语音段的开始和结束分别对应语音段的前端点和后端点，前端点就是非语音段到语音段的边界帧，后端点就是语音段到非语音段的边界帧，因此，语音段是连续一定长度的帧数据都满足语音的要求得到的。

[0074] 与上述语音交互方法相对应，本发明的语音交互系统如图2所示，包括录音模块1、端点检测模块2、前缀词检测模块3、语音活动检测模块4、语音识别模块5、判断模块6、执行模块7，该录音模块1用于录制用户输入的音频数据；该端点检测模块2用于对所述音频数据进行端点检测，直至检测到语音前端点；该前缀词检测模块3用于对从所述语音前端点起的音频数据进行前缀词检测，直至检测到前缀词语音，其中，所述前缀词为反映执行动作类型的词语；该语音活动检测模块4用于获取所述音频数据中从所述前缀词语音的前端点起的语音段作为语音指令，直至检测到指令获取终止事件；该语音识别模块5用于对所述语音指令进行语音识别，得到语音识别结果；该判断模块6用于判断所述语音识别结果是否有效；该执行模块7用于执行有效的语音识别结果。

[0075] 本发明的系统还可进一步包括降噪模块(图中未示出)，该降噪模块分别与录音模块1及端点检测模块2连接，用于对录音模块1录制的音频数据进行降噪处理，并将降噪处理后的音频数据传送给端点检测模块2。

[0076] 进一步地，上述前缀词检测模块3还可用于基于包括前缀词模型和垃圾模型的并行搜索网络，检测从所述语音前端点起的音频数据中是否存在所述前缀词语音。

[0077] 进一步地，上述判断模块6还可用于判断命令词网络中是否存在与所述语音识别结果相匹配的命令词，如存在，则判定所述语音识别结果有效。

[0078] 上述指令获取终止事件例如可包括语音段结束和语音段已持续设定时间，对此，上述端点检测模块2还可用于检测该语音段的后端点及该语音段的持续时间。

[0079] 本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述得比较简单，相关之处参见方法实施例的部分说明即可。以上所描述的系统实施例仅仅是示意性的，其中所述作为分离部件说明的模块或单元可以是或者也可以不是物理上分开的，作为模块或单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。

[0080] 以上依据图式所示的实施例详细说明了本发明的构造、特征及作用效果，以上所述仅为本发明的较佳实施例，但本发明不以图面所示限定实施范围，凡是依照本发明的构想所作的改变，或修改为等同变化的等效实施例，仍未超出说明书与图示所涵盖的精神时，均应在本发明的保护范围内。

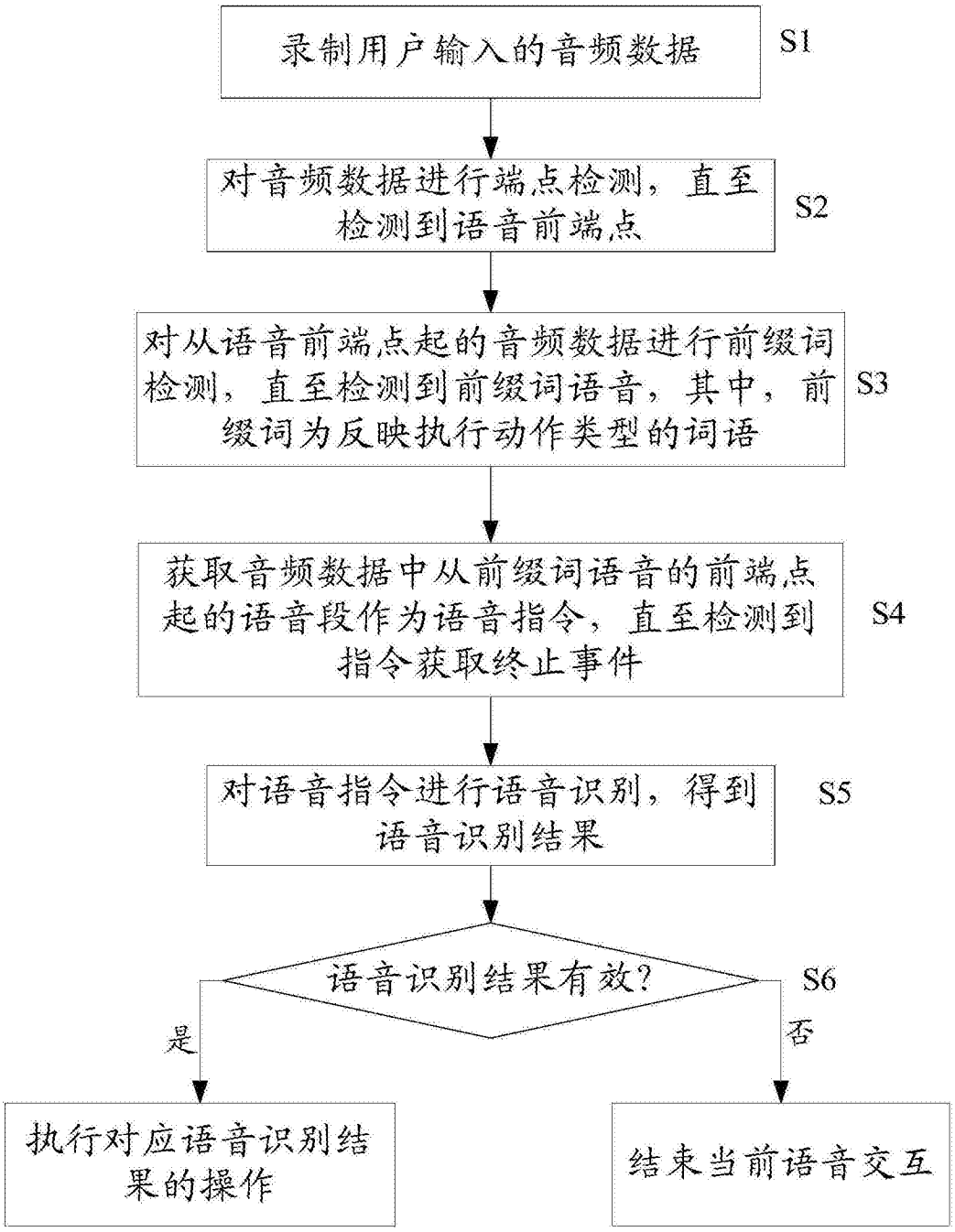

录制用户输入的音频数据
S1
对音频数据进行端点检测，直至检测到语音前端点
S2
对从语音前端点起的音频数据进行前缀词检测，直至检测到前缀词语音，其中，前缀词为反映执行动作类型的词语
S3
获取音频数据中从前缀词语音的前端点起的语音段作为语音指令，直至检测到指令获取终止事件
S4
对语音指令进行语音识别，得到语音识别结果
S5
语音识别结果有效?
S6
是
否
执行对应语音识别结果的操作
结束当前语音交互

图1

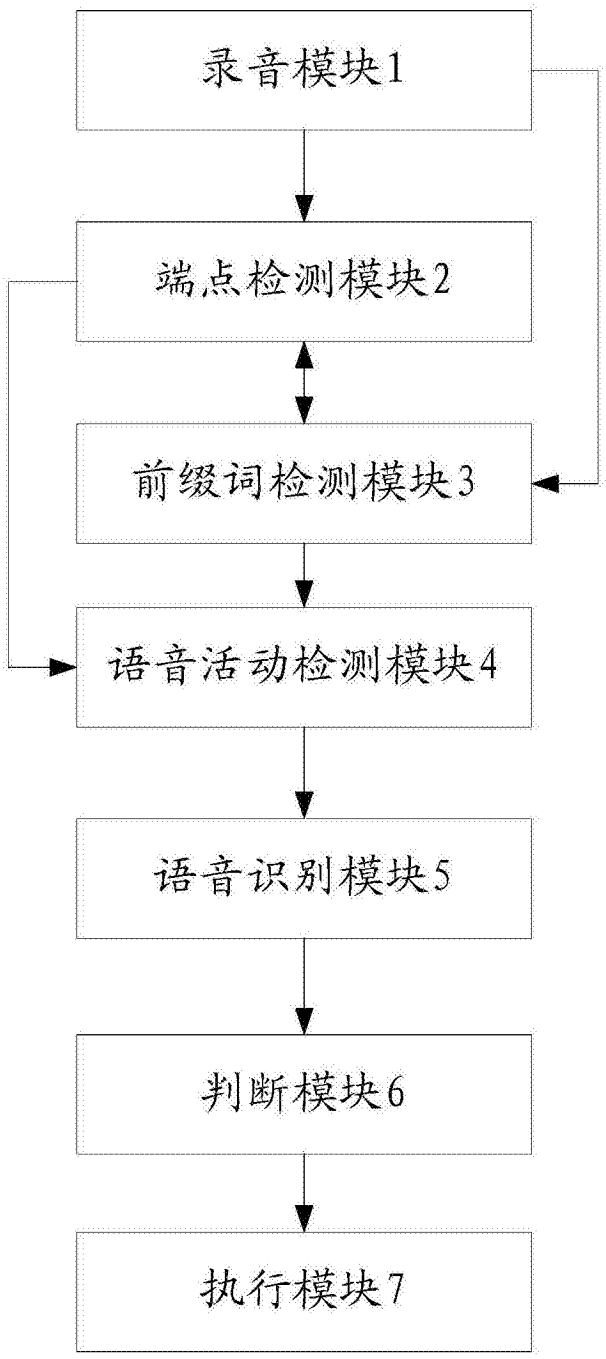


图2