(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 104143329 B
(45)授权公告日 2015.10.21

(21)申请号 201310361835.5
(22)申请日 2013.08.19
(73)专利权人 腾讯科技(深圳)有限公司
地址 518044 广东省深圳市福田区振兴路赛格科技园2栋东403室
(72)发明人 马建雄 李露 卢鲤 张翔 岳帅 饶丰 王尔玉 孔令挥
(74)专利代理机构 北京德琦知识产权代理有限公司 11018
代理人 周华霞 王丽琴
(51)Int.Cl.
G10L 15/08(2006.01)

(56)对比文件
CN 101231660 A,2008.07.30,
CN 101645269 A,2010.02.10,
US 2003220784 A1,2003.11.27,
US 5689616 A,1997.11.18,
US 5805771 A,1998.09.08,

审查员 祝晔

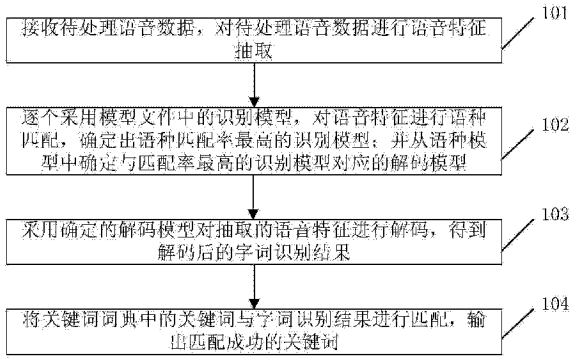
权利要求书2页 说明书6页 附图2页

(54)发明名称

进行语音关键词检索的方法及装置

(57)摘要

本发明公开了进行语音关键词检索的方法及装置，其中，该方法在模型文件中配置至少两类语种模型，每类语种模型包含识别模型及对应的解码模型；该方法包括：接收待处理语音数据，对待处理语音数据进行语音特征抽取；逐个采用模型文件中的识别模型，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型；采用确定的解码模型对抽取的语音特征进行解码，得到解码后的字词识别结果；将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词。本发明方案能够支持至少两种语言的关键词检索，节省成本。



CN 104143329 B

1. 一种进行语音关键词检索的方法，其特征在于，在模型文件中配置至少两类语种模型，每类语种模型包含识别模型及对应的解码模型；该方法包括：

接收待处理语音数据，对待处理语音数据进行语音特征抽取；

逐个采用模型文件中的识别模型，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型；

采用确定的解码模型对抽取的语音特征进行解码，得到解码后的字词识别结果；

将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词，实现在一个检索方案中对两类以上的语种进行关键词检索。

2. 如权利要求 1 所述的方法，其特征在于，当需要进行语种扩展时，该方法还包括：

训练创建新的识别模型和解码模型；

在模型文件中增加语种模型，包含创建的识别模型及对应的解码模型。

3. 如权利要求 1 所述的方法，其特征在于，所述对待处理语音数据进行语音特征抽取包括：

对待处理语音数据进行语音波形处理，从语音波形中提取随时间变化的语音特征序列，提取的语音特征具有区分性。

4. 如权利要求 1、2 或 3 所述的方法，其特征在于，所述采用确定的解码模型对抽取的语音特征进行解码，包括：

采用确定的解码模型对抽取的每一帧语音特征在搜索网络中搜索最佳匹配路径，得到词网，作为解码后的字词识别结果；所述词网包含开始节点和结束节点，以及开始节点和结束节点之间的中间节点，每个节点代表一个时间段对应的词。

5. 如权利要求 4 所述的方法，其特征在于，所述将关键词词典中的关键词与字词识别结果进行匹配，包括：

将最佳匹配路径的词网进行最小错误的对齐操作，生成混淆网络，所述混淆网络按照时间进行排序，给出每个时间段的字词识别结果及字词识别结果的概率；

将关键词词典中的关键词对混淆网络中的各字词识别结果进行匹配，确定出匹配成功的字词识别结果，作为匹配成功的关键词。

6. 一种进行语音关键词检索的装置，其特征在于，该装置包括模型文件配置单元、特征抽取单元、语种识别单元、解码单元和关键词搜索单元；

所述模型文件配置单元，在模型文件中配置至少两类语种模型，每类语种模型包含识别模型及对应的解码模型；

所述特征抽取单元，接收待处理语音数据，对待处理语音数据进行语音特征抽取，将抽取的语音特征发送给所述语种识别单元；

所述语种识别单元，逐个采用模型文件中的识别模型，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型，将抽取的语音特征发送给解码单元；

所述解码单元，采用确定的解码模型对抽取的语音特征进行解码，得到解码后的字词识别结果，发送给所述关键词搜索单元；

所述关键词搜索单元，将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词，实现在一个检索方案中对两类以上的语种进行关键词检索。

7. 如权利要求 6 所述的装置，其特征在于，该装置还包括语种扩展单元，训练创建新的识别模型和解码模型，在模型文件中增加语种模型，包含创建的识别模型及对应的解码模型。

8. 如权利要求 6 所述的装置，其特征在于，所述特征抽取单元包括特征抽取模块，对待处理语音数据进行语音波形处理，从语音波形中提取随时间变化的语音特征序列，提取的语音特征具有区分性。

9. 如权利要求 6、7 或 8 所述的装置，其特征在于，所述解码单元包括路径搜索模块，对每一帧语音特征在搜索网络中搜索最佳匹配路径，得到词网，作为解码后的字词识别结果；所述词网包含开始节点和结束节点，以及开始节点和结束节点之间的中间节点，每个节点代表一个时间段对应的词。

10. 如权利要求 9 所述的装置，其特征在于，所述关键词搜索单元包括混淆网络生成模块和关键词匹配模块；

所述混淆网络生成模块，将最佳匹配路径的词网进行最小错误的对齐操作，生成混淆网络，所述混淆网络按照时间进行排序，给出每个时间段的字词识别结果及字词识别结果的概率；

所述关键词匹配模块，将关键词词典中的关键词对混淆网络中的各字词识别结果进行匹配，确定出匹配成功的字词识别结果，作为匹配成功的关键词。

进行语音关键词检索的方法及装置

技术领域

[0001] 本发明涉及信息处理技术，尤其涉及进行语音关键词检索的方法及装置。

背景技术

[0002] 语音识别技术中，常需要对一段语音进行检索，以确定其是否包含关注的关键词。例如，对会议录音，需要确定其是否为关于计算机的会议，通过检索录音中是否包含“显示器”、“键盘”等关键词进行确定。

[0003] 语音关键词检测的应用现在越来越广泛，但大部分都是针对普通话或者其他特定的某一方言进行，局限性较大。现有语音关键词检索方案中，只针对某一类语种进行关键词检索，将针对该语种的检索算法与语种模型融合在一起，检测算法负责整个检索过程，其中会调用语种模型进行语种识别和解码，解码后，将判别解码结果中是否有关注的关键词，如果有，则输出相应的关键词；如果语音数据不属于该语种，则无法进行识别，需要采用能识别相应语种的另一检测算法对其重新进行关键词检索。

[0004] 综上，现有技术中，语音关键词检索方案只支持某一特定语种的处理，每类语种分别有各自完整的语音关键词检索方案，其局限性很大，且成本较高。

发明内容

[0005] 本发明提供了一种进行语音关键词检索的方法及装置，该方法能够支持至少两种语言的关键词检索，节省成本。

[0006] 本发明提供了一种进行语音关键词检索的方法及装置，该装置能够支持至少两种语言的关键词检索，节省成本。

[0007] 一种进行语音关键词检索的方法，该方法在模型文件中配置至少两类语种模型，每类语种模型包含识别模型及对应的解码模型；该方法包括：

[0008] 接收待处理语音数据，对待处理语音数据进行语音特征抽取；

[0009] 逐个采用模型文件中的识别模型，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型；

[0010] 采用确定的解码模型对抽取的语音特征进行解码，得到解码后的字词识别结果；

[0011] 将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词。

[0012] 较佳地，当需要进行语种扩展时，该方法还包括：

[0013] 训练创建新的识别模型和解码模型；

[0014] 在模型文件中增加语种模型，包含创建的识别模型及对应的解码模型。

[0015] 较佳地，所述对待处理语音数据进行语音特征抽取包括：

[0016] 对待处理语音数据进行语音波形处理，从语音波形中提取随时间变化的语音特征序列，提取的语音特征具有区分性。

[0017] 较佳地，所述采用确定的解码模型对抽取的语音特征进行解码，包括：

[0018] 采用确定的解码模型对抽取的每一帧语音特征在搜索网络中搜索最佳匹配路径，

得到词网，作为解码后的字词识别结果；所述词网包含开始节点和结束节点，以及开始节点和结束节点之间的中间节点，每个节点代表一个时间段对应的词。

[0019] 较佳地，所述将关键词词典中的关键词与字词识别结果进行匹配，包括：

[0020] 将最佳匹配路径的词网进行最小错误的对齐操作，生成混淆网络，所述混淆网络按照时间进行排序，给出每个时间段的字词识别结果及字词识别结果的概率；

[0021] 将关键词词典中的关键词对混淆网络中的各字词识别结果进行匹配，确定出匹配成功的字词识别结果，作为匹配成功的关键词。一种进行语音关键词检索的装置，该装置包括模型文件配置单元、特征抽取单元、语种识别单元、解码单元和关键词搜索单元；

[0022] 所述模型文件配置单元，在模型文件中配置至少两类语种模型，每类语种模型包含识别模型及对应的解码模型；

[0023] 所述特征抽取单元，接收待处理语音数据，对待处理语音数据进行语音特征抽取，将抽取的语音特征发送给所述语种识别单元；

[0024] 所述语种识别单元，逐个采用模型文件中的识别模型，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型，将抽取的语音特征发送给解码单元；

[0025] 所述解码单元，采用确定的解码模型对抽取的语音特征进行解码，得到解码后的字词识别结果，发送给所述关键词搜索单元；

[0026] 所述关键词搜索单元，将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词。

[0027] 较佳地，该装置还包括语种扩展单元，训练创建新的识别模型和解码模型，在模型文件中增加语种模型，包含创建的识别模型及对应的解码模型。

[0028] 较佳地，所述特征抽取单元包括特征抽取模块，对待处理语音数据进行语音波形处理，从语音波形中提取随时间变化的语音特征序列，提取的语音特征具有区分性。

[0029] 较佳地，所述解码单元包括路径搜索模块，对每一帧语音特征在搜索网络中搜索最佳匹配路径，得到词网，作为解码后的字词识别结果；所述词网包含开始节点和结束节点，以及开始节点和结束节点之间的中间节点，每个节点代表一个时间段对应的词。

[0030] 较佳地，所述关键词搜索单元包括混淆网络生成模块和关键词匹配模块；

[0031] 所述混淆网络生成模块，将最佳匹配路径的词网进行最小错误的对齐操作，生成混淆网络，所述混淆网络按照时间进行排序，给出每个时间段的字词识别结果及字词识别结果的概率；

[0032] 所述关键词匹配模块，将关键词词典中的关键词对混淆网络中的各字词识别结果进行匹配，确定出匹配成功的字词识别结果，作为匹配成功的关键词。

[0033] 从上述方案可以看出，本发明中，在模型文件中配置至少两类语种模型，每类语种模型包含识别模型及对应的解码模型；当需要进行关键词检索时，对待处理语音数据进行语音特征抽取；逐个采用模型文件中的识别模型，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型，进行解码后得到解码后的字词识别结果；将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词。采用本发明方案，根据实际需要，可以在模型文件中配置至少两类语种模型，实现在一个检索方案中对两类以上的语种进行关键词检索，从而，解决

了现有技术只支持针对某一特定语种进行处理的缺陷，并且，节省了成本。

附图说明

[0034] 图 1 为本发明进行语音关键词检索的方法示意性流程图；

[0035] 图 2 为本发明进行语音关键词检索的方法流程图实例；

[0036] 图 3 为本发明进行语音关键词检索的装置结构示意图。

具体实施方式

[0037] 为使本发明的目的、技术方案和优点更加清楚明白，下面结合实施例和附图，对本发明进一步详细说明。

[0038] 本发明设置模型文件，在模型文件中配置至少两类语种模型，并基于模型文件进行语音关键词检索，以实现在一个检索方案中对两类以上语种进行处理。

[0039] 参见图 1，为本发明进行语音关键词检索的方法示意性流程图，该方法预先设置模型文件，在模型文件中配置至少两类语种模型，每类语种模型包含识别模型及对应的解码模型；每个识别模型对某一特征语种的语音进行识别，确定为本识别模型支持的语种后，发送给与本识别模型对应的解码模型进行解码。

[0040] 图 1 的流程包括以下步骤：

[0041] 步骤 101，接收待处理语音数据，对待处理语音数据进行语音特征抽取。

[0042] 实现时，本步骤具体包括：对待处理语音数据进行语音波形处理，从语音波形中提取随时间变化的语音特征序列，提取的语音特征具有区分性。

[0043] 步骤 102，逐个采用模型文件中的识别模型，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型。

[0044] 识别模型用于对语音进行语种识别，以确定是否为本识别模型能够识别的语种。

[0045] 步骤 103，采用确定的解码模型对抽取的语音特征进行解码，得到解码后的字词识别结果。

[0046] 实现时，本步骤可具体包括：采用确定的解码模型对抽取的每一帧语音特征在搜索网络中搜索最佳匹配路径，得到最可能的识别结果，作为解码后的识别结果，识别结果为至少一个。

[0047] 所述搜索网络具体如加权有限状态转换机 (WFST，Weighted Finite State Transducers) 搜索网络，WFST 搜索网络是一张合成了声学模型、语言模型以及词表的搜索网络，解码模型将依据该 WFST 搜索网络进行解码计算，最终输出经过一定裁剪后的词网，该词网拥有一个开始节点和一个结束节点，以及开始节点和结束节点之间的中间节点，每个节点代表某一时间段可能的词，从开始节点到结束节点之间有至少一条路径，每条路径代表一个识别结果。

[0048] 例如，某实例中，从开始节点到结束节点之间有两条路径，其中一条路径有 5 个节点，从开始节点到结束节点的节点序列对应的词为‘我’，‘们’，‘吃’，‘饭’，‘吧’，也就是识别结果为“我们吃饭吧“；另一条路径也有 5 个节点，从开始节点到结束节点的节点序列对应的词为‘我’，‘们’，‘迟’，‘饭’，‘吧’，也就是，另一种识别结果为“我们迟饭吧”。

[0049] 步骤 104，将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词。

[0050] 如果步骤 103 得到的字词识别结果，是在搜索网络中搜索出的最佳匹配路径；相应地，本步骤体包括：

[0051] 将最佳匹配路径的词网进行最小错误的对齐操作，生成混淆网络，所述混淆网络按照时间进行排序，给出每个时间段的字词识别结果及字词识别结果的概率；将关键词词典中的关键词对混淆网络中的各字词识别结果进行匹配，确定出匹配成功的字词识别结果，作为匹配成功的关键词。

[0052] 最小错误的对齐操作为现有技术，该技术能够对最佳匹配路径的词网进行分析，确定出某一时间段可能对应的多种识别结果，并能给出各字词识别结果的概率。仍然以前述“我们吃饭吧”及“我们迟饭吧”的实例进行说明，采用最小错误的对其操作之后，确定出第 1、2 节点对应的识别结果为‘我’、‘们’；第 3 节点对应的识别结果为‘吃’和‘迟’，并给出为‘吃’、‘迟’的概率；第 4、5 节点对应的识别结果为‘饭’、‘吧’。如果开始节点与结束节点之间只有一条路径，则无需采用最小错误对齐操作进行分析处理。

[0053] 关键词词典中包含了关注的关键词，将关键词词典中的所有关键词分别与各字词识别结果进行匹配，如果相同，则确定为匹配成功的字词识别结果。如果关键词词典中包含“吃饭”、“蔬菜”、“素食”，则针对上述的实例，匹配后输出的关键词为“吃饭”。

[0054] 本发明中，在模型文件中配置至少两类语种模型，每类语种模型包含识别模型及对应的解码模型；当需要进行关键词检索时，对待处理语音数据进行语音特征抽取；逐个采用模型文件中的识别模型，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型，进行解码后得到解码后的字词识别结果；将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词。采用本发明方案，根据实际需要，可以在模型文件中配置至少两类语种模型，实现对两类以上的语种进行关键词检索，从而，解决了现有技术只支持针对某一特定语种进行处理的缺陷，并且，节省了成本。

[0055] 现有语音关键词检索方案中，只针对某一类语种进行关键词检索，具体实现时，将针对该语种的检测算法和语种模型融合在一起，这样处理缺乏可扩展性，即当有其他方言的需求时无法动态支持。采用本发明方案后，当需要进行语种扩展时，训练创建针对该语种的识别模型和解码模型；在模型文件中增加语种模型，增加的语种模型包含创建的识别模型及对应的解码模型。这样，后续便可结合新增的语种模型进行关键词检索。

[0056] 下面通过图 2 的流程对本发明进行语音关键词检索的方法进行实例说明，模型文件中已配置了关于语种 A 和 B 的两类语种模型，每类语种模型包含识别模型及对应的解码模型，该方法包括以下步骤：

[0057] 步骤 201，接收关于语种 C 的扩展指令。

[0058] 步骤 202，训练创建关于语种 C 的识别模型 C 和解码模型 C，在模型文件中增加语种模型 C，其中包含创建的识别模型 C 及解码模型 C。

[0059] 训练关于某语种的识别模型和解码模型，可采用现有方案实现，这里不赘述。

[0060] 步骤 203，接收待处理语音数据，对待处理语音数据进行语音特征抽取。

[0061] 该过程目的是从语音波形中提取随时间变化的语音特征序列，提取的特征参数能

有效地代表语音特征，具有很好的区分性，作为后续处理的基础数据。

[0062] 步骤 204，分别采用模型文件中的识别模型 A、识别模型 B 和识别模型 C，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型。

[0063] 本实例中，假设匹配率最高的为识别模型 C，对应着解码模型 C。识别模型对语音特征的识别，可采用现有方案实现。

[0064] 步骤 205，采用解码模型 C 对抽取的语音特征进行解码，得到解码后的字词识别结果。

[0065] 解码模型，是针对相应语种的语音进行解码过程中使用的模型；解码模型采用声学模型、语言模型以及词表组合而成，可对抽取的语音特征进行解析，生成经过一定裁剪后的词网，后续算法在此搜索网络中进行计算以得到最后的关键词结果。解码模型对语音特征的解码，可采用现有方案实现。

[0066] 步骤 206，将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词。

[0067] 本实例将关键词检索的算法与模型分离，从而使动态扩展方言支持成为可能。在需要支持新的方言时，只需要针对新的方言训练新的模型，并进行配置即可支持新的方言关键词检测。相比现有将检索算法与语种模型高度融合的方案，其扩展性是其最大的特点，可以根据实际需求灵活增加或者取消对特定语种的支持，也降低了因需求而不断升级的成本。另外可维护性也具有一定的优势，将检测算法与语种模型分离是两个部分功能明确，结构更加清晰，部署相对也简单。

[0068] 参见图 3，为本发明进行语音关键词检索的装置结构示意图，该装置包括模型文件配置单元、特征抽取单元、语种识别单元、解码单元和关键词搜索单元；

[0069] 所述模型文件配置单元，在模型文件中配置至少两类语种模型，每类语种模型包含识别模型及对应的解码模型；

[0070] 所述特征抽取单元，接收待处理语音数据，对待处理语音数据进行语音特征抽取，将抽取的语音特征发送给所述语种识别单元；

[0071] 所述语种识别单元，逐个采用模型文件中的识别模型，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型，将抽取的语音特征发送给解码单元；

[0072] 所述解码单元，采用确定的解码模型对抽取的语音特征进行解码，得到解码后的字词识别结果，发送给所述关键词搜索单元；

[0073] 所述关键词搜索单元，将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词。

[0074] 较佳地，该装置还包括语种扩展单元，训练创建新的识别模型和解码模型，在模型文件中增加语种模型，包含创建的识别模型及对应的解码模型。

[0075] 较佳地，所述特征抽取单元包括特征抽取模块，对待处理语音数据进行语音波形处理，从语音波形中提取随时间变化的语音特征序列，提取的语音特征具有区分性。

[0076] 较佳地，所述解码单元包括路径搜索模块，采用确定的解码模型对抽取的每一帧语音特征在搜索网络中搜索最佳匹配路径，得到词网，作为解码后的字词识别结果；所述词

网包含开始节点和结束节点，以及开始节点和结束节点之间的中间节点，每个节点代表一个时间段对应的词。

[0077] 较佳地，所述关键词搜索单元包括混淆网络生成模块和关键词匹配模块；

[0078] 所述混淆网络生成模块，将最佳匹配路径的词网进行最小错误的对齐操作，生成混淆网络，所述混淆网络按照时间进行排序，给出每个时间段的字词识别结果及字词识别结果的概率；

[0079] 所述关键词匹配模块，将关键词词典中的关键词对混淆网络中的各字词识别结果进行匹配，确定出匹配成功的字词识别结果，作为匹配成功的关键词。

[0080] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

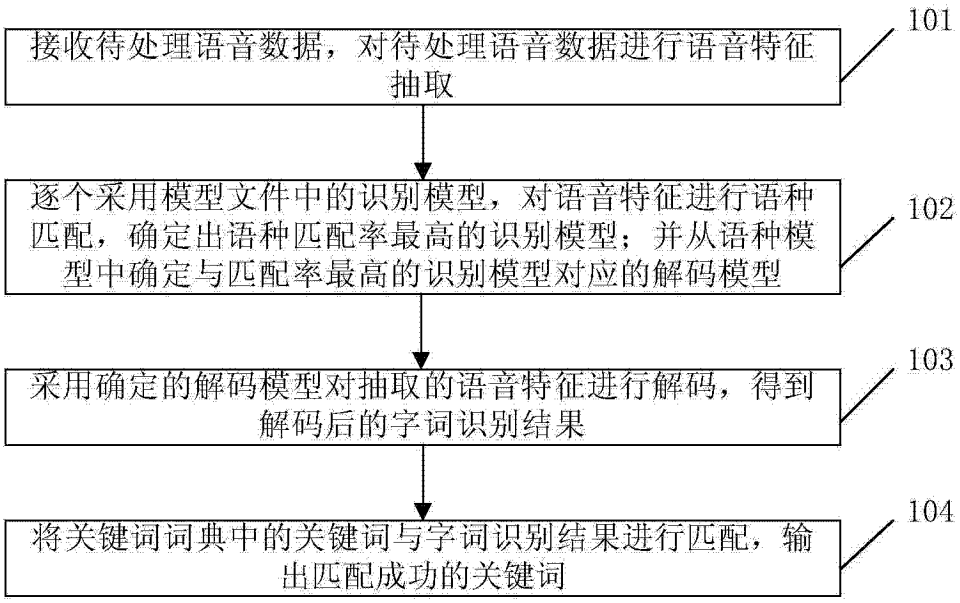

接收待处理语音数据，对待处理语音数据进行语音特征抽取
101
逐个采用模型文件中的识别模型，对语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型
102
采用确定的解码模型对抽取的语音特征进行解码，得到解码后的字词识别结果
103
将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词
104

图 1

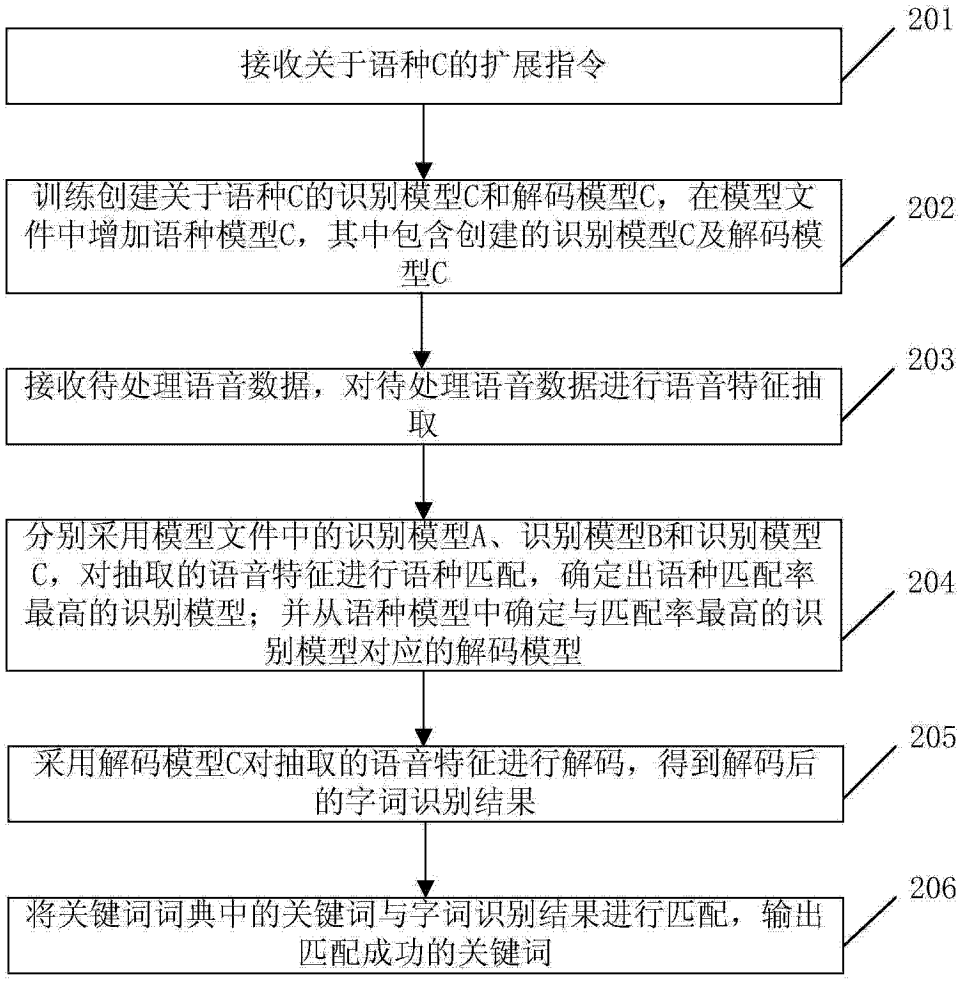

接收关于语种C的扩展指令
201
训练创建关于语种C的识别模型C和解码模型C，在模型文件中增加语种模型C，其中包含创建的识别模型C及解码模型C
202
接收待处理语音数据，对待处理语音数据进行语音特征抽取
203
分别采用模型文件中的识别模型A、识别模型B和识别模型C，对抽取的语音特征进行语种匹配，确定出语种匹配率最高的识别模型；并从语种模型中确定与匹配率最高的识别模型对应的解码模型
204
采用解码模型C对抽取的语音特征进行解码，得到解码后的字词识别结果
205
将关键词词典中的关键词与字词识别结果进行匹配，输出匹配成功的关键词
206

图 2

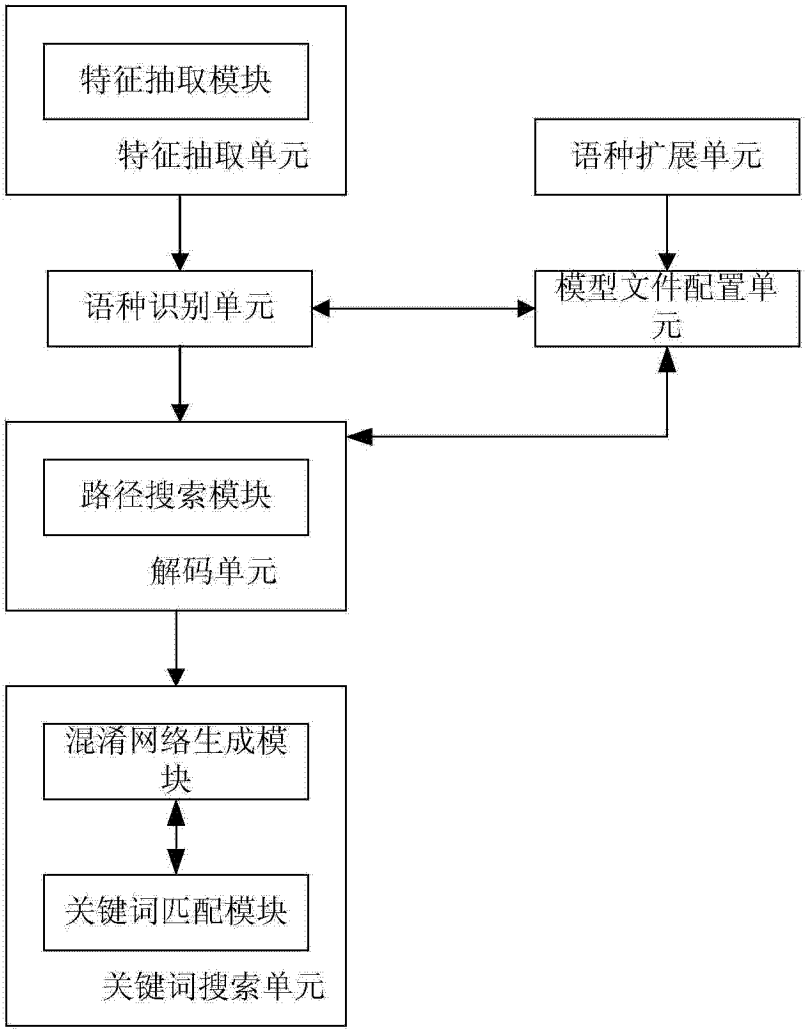

特征抽取模块
特征抽取单元
语种扩展单元
语种识别单元
模型文件配置单元
路径搜索模块
解码单元
混淆网络生成模块
关键词匹配模块
关键词搜索单元

图 3