(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
2017 年 5 月 11 日 (11.05.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/076222 A1

(51) 国际专利分类号:
G10L 15/30 (2013.01) *G06F 17/27* (2006.01)

(21) 国际申请号: PCT/CN2016/103691

(22) 国际申请日: 2016 年 10 月 28 日 (28.10.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201510752397.4 2015 年 11 月 6 日 (06.11.2015) CN

(71) 申请人: **阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED)** [—/CN]; 英属开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人: **李晓辉 (LI, Xiaohui)**; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 **李宏言 (LI, Hongyan)**; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: **北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.)**; 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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, 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, JP, KE, KG, 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))。

(54) Title: SPEECH RECOGNITION METHOD AND APPARATUS

(54) 发明名称 : 语音识别方法及装置

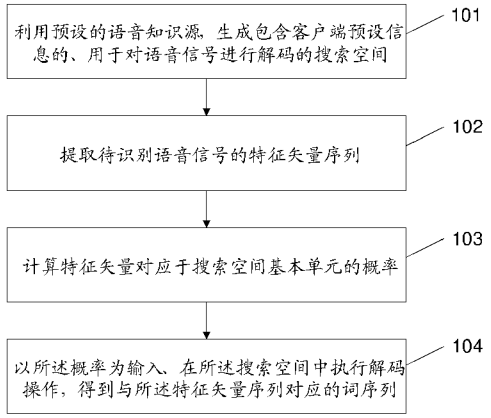

图 1

101 GENERATE SEARCH SPACE THAT COMPRISES PRESET INFORMATION OF A CLIENT AND IS USED FOR DECODING A SPEECH SIGNAL BY UTILIZING A PRESET SPEECH KNOWLEDGE SOURCE
102 EXTRACT A CHARACTERISTIC VECTOR SEQUENCE OF A SPEECH SIGNAL TO BE RECOGNIZED
103 CALCULATE A PROBABILITY AT WHICH A CHARACTERISTIC VECTOR CORRESPONDS TO A BASIC UNIT OF THE SEARCH SPACE
104 EXECUTE A DECODING OPERATION IN THE SEARCH SPACE BY USING THE PROBABILITY AS AN INPUT, TO OBTAIN A WORD SEQUENCE CORRESPONDING TO THE CHARACTERISTIC VECTOR SEQUENCE

(57) **Abstract**: A speech recognition method comprises: generating search space that comprises preset information of a client and is used for decoding a speech signal by utilizing a preset speech knowledge source (101); extracting a characteristic vector sequence of a speech signal to be recognized (102); calculating a probability at which a characteristic vector corresponds to a basic unit of the search space (103); and executing a decoding operation in the search space by using the probability as an input, to obtain a word sequence corresponding to the characteristic vector sequence (104). Also provided are a speech recognition apparatus and another speech recognition method and apparatus. By using the methods, preset information of a client is comprised when search space for use in decoding is generated, and therefore relevant information of the client can be recognized accurately when recognizing a speed signal acquired by the client; and thus, the accuracy of speed recognition can be increased and user experience can be improved.

(57) 摘要:

[见续页]

WO 2017/076222 A1

一种语音识别方法，包括：利用预设的语音知识源，生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间（101）；提取待识别语音信号的特征矢量序列（102）；计算特征矢量对应于搜索空间基本单元的概率（103）；以该概率为输入、在搜索空间中执行解码操作，得到与该特征矢量序列对应的词序列（104）。还提供一种语音识别装置，以及另一种语音识别方法及装置。采用上述方法，由于在生成用于解码的搜索空间时包含了客户端预设信息，因此在对客户端采集的语音信号进行识别时能够相对准确地识别出与客户端相关的信息，从而可以提高语音识别的准确率，提升用户的使用体验。

语音识别方法及装置

本申请要求 2015 年 11 月 06 日递交的申请号为 201510752397.4、发明名称为“语音识别方法及装置”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

本申请涉及语音识别技术，具体涉及一种语音识别方法及装置。本申请同时涉及另一种语音识别方法及装置。

背景技术

语音是语言的声学表现，是人类交流信息最自然、最有效、最方便的手段，也是人类思维的一种依托。自动语音识别（Automatic Speech Recognition—ASR）通常是指让计算机等设备通过对语音的识别和理解，把人的口语转化为相应的输出文本或者命令的过程。其核心框架是：在利用统计模型建模的基础上，根据从待识别语音信号中提取的特征序列 O，利用下述贝叶斯决策准则来求解与待识别语音信号对应的最佳词序列 W*：

$$W^* = \text{argmax}P(O|W)P(W)$$

在具体实施中，上述求解最佳词序列的过程称为解码过程（实现解码功能的模块通常称为解码器），即：在由发音词典、语言模型等多种知识源组成的搜索空间中搜索出上式所示的最佳词序列。

随着技术的发展，硬件的计算能力和存储容量有了很大的进步，语音识别系统已经逐步在业界得以应用，在客户端设备上也出现了各种用语音作为人机交互媒介的应用，例如智能手机上的拨打电话应用，用户只需发出语音指示（如：“给张三打电话”），即可自动实现电话拨打功能。

目前的语音识别应用通常采用两种模式，一种是基于客户端和服务器的模式，即：客户端采集语音，经由网络上传至服务器，服务器通过解码将语音识别为文本，然后回传到客户端。之所以采用这样的模式，是因为客户端的计算能力相对较弱，其内存空间也比较有限，而服务器在这两方面都具有明显的优势；但是采用这种模式，如果没有网络接入环境，客户端则无法完成语音识别功能。针对上述问题出现了仅依赖于客户端的第二种语音识别应用模式，在该模式下，通过缩减规模，将原本存放在服务器上的模型和搜索空间放在客户端设备本地，由客户端自行完成采集语音以及解码的操作。

在实际应用中，无论是第一种模式还是第二种模式，在采用上述通用框架进行语音

识别时，通常无法有效识别语音信号中与客户端设备本地信息相关的内容，例如：通讯录中的联系人名称，从而导致识别准确率低，给用户的使用带来不便，影响用户的使用体验。

发明内容

本申请实施例提供一种语音识别方法和装置，以解决现有的语音识别技术对客户端本地相关信息的识别准确率低的问题。本申请实施例还提供另一种语音识别方法和装置。

本申请提供一种语音识别方法，包括：

利用预设的语音知识源，生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间；

提取待识别语音信号的特征矢量序列；

计算特征矢量对应于搜索空间基本单元的概率；

以所述概率为输入、在所述搜索空间中执行解码操作，得到与所述特征矢量序列对应的词序列。

可选的，所述搜索空间包括：加权有限状态转换器。

可选的，所述搜索空间基本单元包括：上下文相关的三音素；

所述预设的知识源包括：发音词典、语言模型、以及三音素状态绑定列表。

可选的，所述利用预设的语音知识源生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间，包括：

通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息，并得到基于三音素状态绑定列表、发音词典以及语言模型的单一加权有限状态转换器；

其中，所述语言模型是采用如下方式预先训练得到的：

将用于训练语言模型的文本中的预设命名实体替换为与预设主题类别对应的标签，并利用所述文本训练语言模型。

可选的，所述通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息，并得到基于三音素状态绑定列表、发音词典以及语言模型的单一加权有限状态转换器，包括：

通过替换标签的方式，向预先生成的基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息；

将添加了客户端预设信息的所述加权有限状态转换器、与预先生成的基于三音素状态绑定列表和发音词典的加权有限状态转换器进行合并，得到所述单一加权有限状态转换器。

可选的，所述用于训练语言模型的文本是指，针对所述预设主题类别的文本。

可选的，所述预设主题类别的数目至少为 2 个；所述语言模型的数目、以及所述至少基于语言模型的加权有限状态器的数目分别与预设主题类别的数目一致；

所述通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息，包括：

确定待识别语音信号所属的预设主题类别；

选择预先生成的、与所述预设主题类别相对应的所述至少基于语言模型的加权有限状态转换器；

通过用与所述预设主题类别对应的客户端预设信息替换相应标签的方式，向所选的加权有限状态转换器中添加客户端预设信息。

可选的，所述确定待识别语音信号所属的预设主题类别，采用如下方式实现：

根据采集所述语音信号的客户端类型、或应用程序确定所述所属的预设主题类别。

可选的，所述预设主题类别包括：拨打电话或发送短信，播放乐曲，或者，设置指令；

相应的客户端预设信息包括：通讯录中的联系人名称，曲库中的乐曲名称，或者，指令集中的指令。

可选的，所述合并操作包括：采用基于预测的方法进行合并。

可选的，预先训练所述语言模型所采用的词表与所述发音词典包含的词一致。

可选的，所述计算特征矢量对应于搜索空间基本单元的概率，包括：

采用预先训练的 DNN 模型计算特征矢量对应于各三音素状态的概率；

根据特征矢量对应于所述各三音素状态的概率，采用预先训练的 HMM 模型计算特征矢量对应于各三音素的概率。

可选的，通过如下方式提升所述采用预先训练的 DNN 模型计算特征矢量对应于各三音素状态的概率的步骤的执行速度：利用硬件平台提供的数据并行处理能力。

可选的，所述提取待识别语音信号的特征矢量序列，包括：

按照预先设定的帧长度对待识别语音信号进行分帧处理，得到多个音频帧；

提取各音频帧的特征矢量，得到所述特征矢量序列。

可选的，所述提取各音频帧的特征矢量包括：提取 MFCC 特征、PLP 特征、或者 LPC 特征。

可选的，在所述得到与所述特征矢量序列对应的词序列后，执行下述操作：

通过与所述客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

可选的，所述通过与所述客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果得到相应的语音识别结果，包括：

从所述词序列中选择对应于所述客户端预设信息的待验证词；

在所述客户端预设信息中查找所述待验证词；

若找到，则判定通过所述准确性验证，并将所述词序列作为语音识别结果；否则通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

可选的，所述通过基于拼音的模糊匹配方式修正所述词序列，包括：

将所述待验证词转换为待验证拼音序列；

将所述客户端预设信息中的各个词分别转换为比对拼音序列；

依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；

用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

可选的，所述相似度包括：基于编辑距离计算的相似度。

可选的，所述方法在客户端设备上实施；所述客户端设备包括：智能移动终端、智能音箱、或者机器人。

相应的，本申请还提供一种语音识别装置，包括：

搜索空间生成单元，用于利用预设的语音知识源，生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间；

特征矢量提取单元，用于提取待识别语音信号的特征矢量序列；

概率计算单元，用于计算特征矢量对应于搜索空间基本单元的概率；

解码搜索单元，用于以所述概率为输入、在所述搜索空间中执行解码操作，得到与所述特征矢量序列对应的词序列。

可选的，所述搜索空间生成单元具体用于，通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息，并得到基于三音素状态绑定列表、发音词典、以及语言模型的单一加权有限状态转换器；

所述语言模型是由语言模型训练单元预先生成的，所述语言模型训练单元用于，将用于训练语言模型的文本中的预设命名实体替换为与预设主题类别对应的标签，并利用所述文本训练语言模型。

可选的，所述搜索空间生成单元包括：

第一客户端信息添加子单元，用于通过替换标签的方式，向预先生成的基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息；

加权有限状态转换器合并子单元，用于将添加了所述客户端预设信息的所述加权有限状态转换器、与预先生成的基于三音素状态绑定列表和发音词典的加权有限状态转换器进行合并，得到所述单一加权有限状态转换器。

可选的，所述解码空间生成单元包括：

第二客户端信息添加子单元，用于通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息；

统一加权有限状态转换器获取子单元，用于在所述第二客户端信息添加子单元完成添加操作之后，得到基于三音素状态绑定列表、发音词典、以及语言模型的单一加权有限状态转换器；

其中，所述第二客户端信息添加子单元包括：

主题确定子单元，用于确定待识别语音信号所属的预设主题类别；

加权有限状态转换器选择子单元，用于选择预先生成的、与所述预设主题类别相对应的所述至少基于语言模型的加权有限状态转换器；

标签替换子单元，用于通过用与所述预设主题类别对应的客户端预设信息替换相应标签的方式，向所选的加权有限状态转换器中添加客户端预设信息。

可选的，所述主题确定子单元具体用于，根据采集所述语音信号的客户端类型、或应用程序确定所述所属的预设主题类别。

可选的，所述加权有限状态转换器合并子单元具体用于，采用基于预测的方法执行合并操作，并得到所述单一加权有限状态转换器。

可选的，所述概率计算单元包括：

三音素状态概率计算子单元，用于采用预先训练的 DNN 模型计算特征矢量对应于各三音素状态的概率；

三音素概率计算子单元，用于根据特征矢量对应于所述各三音素状态的概率，采用预先训练的 HMM 模型计算特征矢量对应于各三音素的概率。

可选的，所述特征矢量提取单元包括：

分帧子单元，用于按照预先设定的帧长度对待识别语音信号进行分帧处理，得到多个音频帧；

特征提取子单元，用于提取各音频帧的特征矢量，得到所述特征矢量序列。

可选的，所述装置包括：

准确性验证单元，用于在所述解码搜索单元得到与特征矢量序列对应的词序列后，通过与所述客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

可选的，所述准确性验证单元包括：

待验证词选择子单元，用于从所述词序列中选择对应于所述客户端预设信息的待验证词；

查找子单元，用于在所述客户端预设信息中查找所述待验证词；

识别结果确认子单元，用于当所述查找子单元找到所述待验证词之后，判定通过所述准确性验证，并将所述词序列作为语音识别结果；

识别结果修正子单元，用于当所述查找子单元未找到所述待验证词之后，通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

可选的，所述识别结果修正子单元，包括：

待验证拼音序列转换子单元，用于将所述待验证词转换为待验证拼音序列；

比对拼音序列转换子单元，用于将所述客户端预设信息中的各个词分别转换为比对拼音序列；

相似度计算选择子单元，用于依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；

待验证词替换子单元，用于用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

此外，本申请还提供另一种语音识别方法，包括：

通过解码获取与待识别语音信号对应的词序列；

通过与客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

可选的，所述通过与客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果，包括：

从所述词序列中选择对应于所述客户端预设信息的待验证词；

在所述客户端预设信息中查找所述待验证词；

若找到，则判定通过所述准确性验证，并将所述词序列作为语音识别结果；否则通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

可选的，所述通过基于拼音的模糊匹配方式修正所述词序列，包括：

将所述待验证词转换为待验证拼音序列；

将所述客户端预设信息中的各个词分别转换为比对拼音序列；

依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；

用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

相应的，本申请还提供另一种语音识别装置，包括：

词序列获取单元，用于通过解码获取与待识别语音信号对应的词序列；

词序列验证单元，用于通过与客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

可选的，所述词序列验证单元包括：

待验证词选择子单元，用于从所述词序列中选择对应于所述客户端预设信息的待验证词；

查找子单元，用于在所述客户端预设信息中查找所述待验证词；

识别结果确认子单元，用于当所述查找子单元找到所述待验证词之后，判定通过所述准确性验证，并将所述词序列作为语音识别结果；

识别结果修正子单元，用于当所述查找子单元未找到所述待验证词之后，通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

可选的，所述识别结果修正子单元，包括：

待验证拼音序列转换子单元，用于将所述待验证词转换为待验证拼音序列；

比对拼音序列转换子单元，用于将所述客户端预设信息中的各个词分别转换为比对拼音序列；

相似度计算选择子单元，用于依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；

待验证词替换子单元，用于用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

与现有技术相比，本申请具有以下优点：

本申请提供的语音识别方法，在利用预设的语音知识源生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间的基础上，计算从待识别语音信号中提取的特征矢量对应于搜索空间基本单元的概率，并且根据所述概率在搜索空间中执行解码操作，从而得到与所述待识别语音信号对应的词序列。本申请提供的上述方法，由于在生成用于解码的搜索空间时包含了客户端预设信息，因此在对客户端采集的语音信号进行识别时能够相对准确地识别出与客户端相关的信息，从而可以提高语音识别的准确率，提升用户的使用体验。

附图说明

图 1 是本申请的一种语音识别方法的实施例的流程图；

图 2 是本申请实施例提供的生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间的处理流程图；

图 3 是本申请实施例提供的执行替换操作前的 G 结构 WFST 的示意图；

图 4 是本申请实施例提供的执行替换操作后的 G 结构 WFST 的示意图；

图 5 是本申请实施例提供的提取待识别语音信号的特征矢量序列的处理流程图；

图 6 是本申请实施例提供的计算特征矢量对应于各三音素的概率的处理流程图；

图 7 是本申请实施例提供的通过文字匹配验证词序列的准确性、并根据验证结果生成相应语音识别结果的处理流程图；

图 8 为本申请实施例提供的语音识别的整体框架图；

图 9 是本申请的一种语音识别装置的实施例的示意图；

图 10 是本申请的另一种语音识别方法的实施例的流程图；

图 11 是本申请的另一种语音识别装置的实施例的示意图。

具体实施方式

在下面的描述中阐述了很多具体细节以便于充分理解本申请。但是，本申请能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本申请内涵的情况下做类似推广，因此，本申请不受下面公开的具体实施的限制。

在本申请中，分别提供了一种语音识别方法及相应装置，以及另一种语音识别方法及相应装置，在下面的实施例中逐一进行详细说明。为了便于理解，在对实施例进行描

述之前，先对本申请的技术方案、相关的技术术语、以及实施例的撰写方式作简要说明。

本申请提供的语音识别方法，通常应用于以语音作为人机交互媒介的应用中，此类应用将采集的语音信号识别为文本，再根据文本执行相应的操作，所述语音信号中通常涉及客户端本地的预设信息（例如，通讯录中的联系人名称）。现有的语音识别技术，对于上述待识别语音信号采用通用的搜索空间进行解码识别，而通用的搜索空间并没有考虑此类应用在不同客户端之间的差异性，因此通常无法有效识别语音信号中与客户端本地信息相关的内容，导致识别准确率低。针对这一问题，本申请的技术方案通过在构建用于对语音信号进行解码的搜索空间的过程中融入客户端预设信息，相当于针对客户端的具体语音识别需求进行了定制，从而能够有效识别与客户端相关的本地信息，达到提高语音识别准确率的目的。

在语音识别系统中，根据待识别的语音信号得到与其最匹配的词序列的过程叫做解码。而本申请所述的用于对语音信号进行解码的搜索空间是指，由语音识别系统涉及的语音知识源（例如：声学模型、发音词典以及语言模型等）所覆盖的、所有可能的语音识别结果所组成的空间。相应的，解码的过程就是针对待识别语音信号在搜索空间中进行搜索和匹配、得到最佳匹配的词序列的过程。

所述搜索空间的形式可以是多样化的，可以采用各种知识源处于相对独立的不同层面的搜索空间，解码的过程就是逐层计算搜索的过程；也可以采用基于加权有限状态转换器（Weighted Finite State Transducer—简称 WFST）的搜索空间，将各种知识源有机融入到统一的 WFST 网络（也称 WFST 搜索空间）中。考虑到后者便于引入不同的知识源、并且可以提高搜索效率，是本申请技术方案进行语音识别的优选方式，因此在本申请提供的实施例中重点描述基于 WFST 网络的实施方式。

所述 WFST 搜索空间，其核心是利用加权有限状态转换器来模拟语言的文法结构以及相关的声学特性。具体操作方法是：将处于不同层次的知识源分别用 WFST 的形式表示，然后运用 WFST 的特性及合并算法，将上述处于不同层次的 WFST 整合成一个单一的 WFST 网络，构成用于进行语音识别的搜索空间。

WFST 网络的基本单元（即驱动 WFST 进行状态转换的基本输入单元）可以根据具体的需求进行选择。考虑音素上下文对音素发音的影响，为了获得更高的识别准确率，在本申请提供的实施例中采用上下文相关的三音素（Context Dependent triphone，简称三音素或者三音子）作为 WFST 网络的基本单元，相应的构建 WFST 搜索空间的知识源包括：三音素状态绑定列表、发音词典、以及语言模型。

所述三音素状态绑定列表通常包含各三音素彼此之间基于发音特点的绑定关系，通常在以三音素为建模单位训练声学模型时，由于三音素可能的组合方式数目众多，为了减少对训练数据的要求，通常可以基于发音特点、采用决策树聚类方法在最大似然准则下对不同的三音素进行聚类，并使用捆绑技术把具有相同发音特点的三音素绑定到一起以便进行参数共享，从而得到所述三音素状态绑定列表；所述发音词典通常包含音素与词之间的对应关系，是桥接在声学层（物理层）和语义层之间的桥梁，让声学层的内容和语义层的内容耦合关联在一起；所述语言模型则提供了语言结构的相关知识，用于计算词序列在自然语言中出现的概率，在具体实施中通常采用 n 元（n-gram）文法语言模型，具体可以通过统计单词之间相互跟随出现的可能性来建模。

采用基于上述知识源构建的 WFST 网络进行语音识别时，为了驱动 WFST 进行所需的搜索，可以先提取待识别语音信号的特征矢量序列，然后利用预先训练好的模型计算从特征矢量对应于各三音素的概率，并根据所述各三音素的概率，在 WFST 搜索空间中执行解码操作，得到与待识别语音信号对应的词序列。

需要说明的是，在本申请提供的实施例中采用上下文相关的三音素作为 WFST 网络的基本单元，在其他实施方式中，也可以采用其他语音单位作为 WFST 网络的基本单元，例如：单音素、或者三音素状态等。采用不同的基本单元，在构建搜索空间时、以及根据特征矢量计算概率时会有一定的差别，例如以三音素状态作为基本单元，那么在构建 WFST 网络时可以融合基于 HMM 的声学模型，在进行语音识别时，则可以计算特征矢量对应于各三音素状态的概率。上述这些都是具体实施方式的变更，只要在构建搜索空间的过程中包含了客户端预设信息，就同样可以实现本申请的技术方案，就都不偏离本申请的技术核心，也都在本申请的保护范围之内。

下面，对本申请的实施例进行详细说明。请参考图 1，其为本申请的一种语音识别方法的实施例的流程图。所述方法包括步骤 101 至步骤 104，在具体实施时，为了提高执行效率，通常可以在步骤 101 之前完成相关的准备工作（此阶段也可以称作准备阶段），生成基于类的语言模型、预设结构的 WFST 以及用于进行语音识别的声学模型等，从而为步骤 101 的执行做好准备。下面先对准备阶段作详细说明。

在准备阶段可以采用如下方式训练语言模型：将用于训练语言模型的文本中的预设命名实体替换为与预设主题类别对应的标签，并利用所述文本训练语言模型。所述命名实体通常是指文本中具有特定类别的实体，例如：人名、歌曲名、机构名、地名等。

下面以拨打电话应用为例进行说明：预设主题类别为拨打电话，对应的标签为

“$CONTACT”，预设命名实体为人名，那么在预先训练语言模型时，可以将训练文本中的人名替换为对应的标签，比如将“我要打电话给小明”中的“小明”替换为“$CONTACT”，然后得到的训练文本为“我要打电话给$CONTACT”。采用进行上述实体替换之后的文本训练语言模型，得到基于类的语言模型。在训练得到上述语言模型的基础上，还可以预先生成基于语言模型的 WFST，以下简称为 G 结构的 WFST。

优选地，为了缩减语言模型以及对应的 G 结构的 WFST 的规模，在预先训练语言模型时，可以选用针对所述预设主题类别的文本（也可以称为基于类的训练文本）进行训练，例如，预设主题类别为拨打电话，那么针对所述预设主题类别的文本可以包括：我要打电话给小明，给小明打个电话等等。

考虑到客户端设备以及以语音作为人机交互媒介的应用程序的多样性，可以预设两个或者两个以上的主题类别，并针对每种主题类别分别预先训练基于类的语言模型、并构建基于所述语言模型的 G 结构 WFST。

在准备阶段还可以预先构建基于发音词典的 WFST，简称为 L 结构的 WFST，以及基于三音素状态绑定列表的 WFST，简称为 C 结构的 WFST，并采用预设的方式对上述各 WFST 进行有针对性的、选择性地合并操作，例如：可以将 C 结构与 L 结构的 WFST 合并为 CL 结构的 WFST，也可以将 L 结构与 G 结构的 WFST 合并为 LG 结构的 WFST，还可以将 C 结构、L 结构以及 G 结构的 WFST 合并为 CLG 结构的 WFST。本实施例在准备阶段生成了 CL 结构的 WFST 以及 G 结构的 WFST（关于合并操作的说明可以参见步骤 101 中的相关文字）。

此外，在准备阶段还可以预先训练用于进行语音识别的声学模型。在本实施例中，每个三音素用一个 HMM（Hidden Markov Model—隐马尔可夫模型）表征，HMM 的隐含状态为三音素中的一个状态（每个三音素通常包含三个状态），采用 GMM（Gaussian mixture model—高斯混合模型）模型确定 HMM 中每个隐含状态输出各特征矢量的发射概率，以从大量语音数据中提取的特征矢量作为训练样本，采用 Baum-Welch 算法学习 GMM 模型和 HMM 模型的参数，可以得到对应于每个状态的 GMM 模型以及对应于每个三音素的 HMM 模型。在后续步骤 103 中则可以使用预先训练好的 GMM 和 HMM 模型计算特征矢量对应于各三音素的概率。

为了提升语音识别的准确率，本实施例在进行语音识别时用 DNN（Deep Neural Networks—深度神经网络）模型替代了 GMM 模型，相应的，在准备阶段可以预先训练用于根据输入的特征矢量输出对应于各三音素状态概率的 DNN 模型。在具体实施时，

可以在训练 GMM 和 HMM 模型的基础上，通过对训练样本进行强制对齐的方式、为训练样本添加对应于各三音素状态的标签，并用打好标签的训练样本训练得到所述 DNN 模型。

需要说明的是，在具体实施时，由于准备阶段的运算量比较大，对内存以及计算速度的要求相对较高，因此准备阶段的操作通常是在服务器端完成的。为了在没有网络接入环境的情况下依然能够完成语音识别功能，本申请提供的方法通常在客户端设备上实施，因此准备阶段生成的各 WFST 以及用于进行声学概率计算的各模型，可以预先安装到客户端设备中，例如：与应用程序一起打包并安装到客户端。

至此，对本实施例涉及的准备阶段进行了详细说明，下面对本实施例的具体步骤 101 至 104 做详细说明。

步骤 101、利用预设的语音知识源，生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间。

本步骤构建 WFST 搜索空间，为后续的语音识别做好准备。在具体实施时，本步骤通常在用语音作为人机交互媒介的客户端应用程序的启动阶段（也称为初始化阶段）执行，通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息，并得到基于三音素状态绑定列表、发音词典、以及语言模型的单一加权有限状态转换器。

本步骤的处理过程可以包括以下步骤 101-1 至 101-4，下面结合图 2 做进一步说明。

步骤 101-1、确定待识别语音信号所属的预设主题类别。

在具体实施时，可以根据采集所述语音信号的客户端类型、或应用程序确定所述所属的预设主题类别。所述预设主题类别包括：拨打电话、发送短信、播放乐曲、设置指令、或者其他应用场景相关的主题类别。其中，与拨打电话或发送短信对应的客户端预设信息包括：通讯录中的联系人名称；与播放乐曲对应的客户端预设信息包括：曲库中的乐曲名称；与设置指令对应的客户端预设信息包括：指令集中的指令；对于其他应用场景相关的主题类别，也同样可以与应用场景所涉及的客户端预设信息相对应，此处不再一一赘述。

例如：对于智能手机，可以根据客户端类型确定待识别语音信号所属的预设主题类别为：拨打电话或发送短信，相应的客户端预设信息为通讯录中的联系人名称；对于智能音箱，可以确定主题类别为：播放乐曲，相应的客户端预设信息为曲库中的乐曲名称；对于机器人，可以确定主题类别为：设置指令，相应的客户端预设信息为指令集中的指

令。

考虑到客户端设备可以同时具有多个用语音作为人机交互媒介的应用，不同的应用涉及不同的客户端预设信息，例如：智能手机也可以安装基于语音交互的音乐播放器，在这种情况下可以根据当前启动的应用程序确定待识别语音信号所属的预设主题类别。

步骤 101-2、选择预先生成的、与所述预设主题类别相对应的 G 结构 WFST。

对于存在多个预设主题类别的情况，在准备阶段通常会生成多个 G 结构 WFST，每个 G 结构 WFST 分别与不同的预设主题类别相对应。本步骤从预先生成的多个 G 结构 WFST 中选择与步骤 101-1 所确定的预设主题类别相对应的 G 结构 WFST。

步骤 101-3、通过用与所述预设主题类别对应的客户端预设信息替换相应标签的方式，向所选的 G 结构 WFST 中添加客户端预设信息。

在准备阶段针对每种预设主题类别训练基于类的语言模型时，将训练文本中的预设命名实体替换为了与相应主题类别对应的标签，例如主题类别为拨打电话或发送短信，将训练文本中的人名替换为“$CONTACT”标签，主题类别为播放乐曲，将训练文本中的乐曲名称替换为“$SONG”标签，因此，生成的 G 结构的 WFST 中通常包含与预设主题类别对应的标签信息。本步骤用与步骤 101-1 所确定的预设主题类别对应的客户端预设信息，替换步骤 101-2 所选 G 结构 WFST 中的相应标签，从而实现向所选 G 结构 WFST 中添加客户端预设信息的目的。

例如，主题类别为拨打电话或者发送短信，则可以将 G 结构 WFST 中的“$CONTACT”标签替换为客户端本地通讯录中的联系人名称，如“张三”、“李四”等；主题类别为播放乐曲，则可以将 G 结构 WFST 中的“$SONG”标签替换为客户端本地曲库中的歌曲名称，如“义勇军进行曲”等。具体的替换，可以通过将与所述标签对应的状态转移链路替换为若干组并行的状态转移链路的方式实现。请参见图 3 和图 4 给出的用客户端通讯录中的联系人进行替换的例子，其中图 3 为替换前的 G 结构 WFST 的示意图，图 4 为用通讯录中的“张三”和“李四”进行替换后得到的 G 结构 WFST 的示意图。

步骤 101-4、将添加了客户端预设信息的 G 结构的 WFST、与预先生成的 CL 结构的 WFST 进行合并，得到单一的 WFST 网络。

在本实施例中，语音识别所用到的知识源涉及从语言层（语言模型）到物理层（三音素状态绑定列表）的内容，本步骤的任务是将不同层次的 WFST 合并（也称为组合、结合）到一起，得到单一的 WFST 网络。

对于两个 WFST，进行合并的基本条件是：其中一个 WFST 的输出符号是另外一个

WFST 输入符号集合的子集。在满足上述要求的前提下，如果将两个 WFST，例如分别为 A 和 B，整合成一个新的 WFST：C，那么 C 的每个状态都由 A 的状态和 B 的状态组成，C 的每个成功路经，都由 A 的成功路经 Pa 和 B 的成功路径 Pb 组成，输入为 i[P] = i[Pa]，输出为 o[P]=o[Pb]，其加权值为由 Pa 和 Pb 的加权值进行相应运算后得到，最后得到的 C 包含 A 和 B 共有的有限状态转换器特性以及搜索空间。在具体实施时，可以采用 OpenFst 库提供的合并算法完成两个 WFST 的合并操作。

具体到本实施例，可以这样理解，L 结构的 WFST 可以看作是单音素与词之间的对应关系，C 结构的 WFST 则在三音素与单音素之间建立对应关系，其输出和 L 结构 WFST 的输入相互对应，可以进行合并，在本实施例的准备阶段已经通过合并得到了 CL 结构的 WFST，本步骤将所述 CL 结构的 WFST 与步骤 101-3 中添加了客户端预设信息的 G 结构 WFST 进行合并，得到了一个输入为三音素概率，输出为词序列的 WFST 网络，从而将处于不同层次的分别对应不同知识源的 WFST，整合为一个单一的 WFST 网络，构成了用于进行语音识别的搜索空间。

优选地，为了加快 CL 结构 WFST 和 G 结构 WFST 的合并速度，减少初始化的耗时，本实施例在执行所述合并操作时没有采用常规的 WFST 合并方法，而是采用了基于预测的合并方法（lookahead 合并方法）。所述 lookahead 合并方法，即在两个 WFST 的合并过程中，通过对未来路径的预测，判断当前执行的合并操作是否会导致无法到达的最终状态（non-coaccessible state），如果是，则阻塞当前操作、不再执行后续的合并操作。通过预测可以提前终止没有必要的合并操作，不仅可以节省合并时间，而且可以缩减最终生成的 WFST 的规模，减少对存储空间的占用。具体实施时，可以采用 OpenFst 库提供的具备 lookahead 功能的过滤器（filter），实现上述预测筛选功能。

优选地，为了加快 CL 结构 WFST 和 G 结构 WFST 的合并速度，在本实施例中预先训练所述语言模型所采用的词表与所述发音词典包含的词是一致的。一般而言，词表中的词的数目通常大于发音词典中的词的数目，而词表中的词的数目和 G 结构的 WFST 的大小有直接关系，如果 G 结构的 WFST 比较大，和 CL 结构的 WFST 进行合并就比较耗时，所以本实施例在准备阶段训练语言模型时，缩减了词表的规模，让词表中的词与发音词典中的词保持一致，从而达到了缩减 CL 结构 WFST 和 G 结构 WFST 的合并时间的效果。

至此，通过步骤 101-1 至 101-4，完成了本技术方案的初始化过程，生成了包含客户端预设信息的 WFST 搜索空间。

需要说明的是，本实施例在准备阶段预先完成 CL 结构的 WFST 的合并、并生成 G 结构的 WFST，在本步骤 101 中则向 G 结构 WFST 中添加客户端预设信息，并将 CL 结构和 G 结构合并得到单一的 WFST。在其他实施方式中，也可以采用其他合并策略，例如，在准备阶段预先完成 LG 结构的 WFST 的合并，在本步骤 101 中向该 WFST 中添加客户端预设信息，然后再与准备阶段生成的 C 结构 WFST 进行合并；或者，在准备阶段直接完成 CLG 结构的 WFST 的合并，并在本步骤 101 中向该 WFST 中添加客户端预设信息也是可以的。考虑到准备阶段生成的 WFST 要占据客户端的存储空间，在有多个基于类的语言模型（相应有多个 G 结构的 WFST）的应用场景中，如果在准备阶段将每个 G 结构 WFST 与其他 WFST 进行合并，将占据较大存储空间，因此本实施例采取的合并方式是优选实施方式，可以减少在准备阶段生成的 WFST 对客户端存储空间的占用。

步骤 102、提取待识别语音信号的特征矢量序列。

待识别语音信号通常是时域信号，本步骤通过分帧和提取特征矢量两个处理过程，获取能够表征所述语音信号的特征矢量序列，下面结合附图 5 做进一步说明。

步骤 102-1、按照预先设定的帧长度对待识别语音信号进行分帧处理，得到多个音频帧。

在具体实施时，可以根据需求预先设定帧长度，例如可以设置为 10ms、或者 15ms，然后根据所述帧长度对待识别语音信号进行逐帧切分，从而将语音信号切分为多个音频帧。根据所采用的切分策略的不同，相邻音频帧可以不存在交叠、也可以是有交叠的。

步骤 102-2、提取各音频帧的特征矢量，得到所述特征矢量序列。

将待识别语音信号切分为多个音频帧后，可以逐帧提取能够表征所述语音信号的特征矢量。由于语音信号在时域上的描述能力相对较弱，通常可以针对每个音频帧进行傅里叶变换，然后提取频域特征作为音频帧的特征矢量，例如，可以提取 MFCC（Mel Frequency Cepstrum Coefficient—梅尔频率倒谱系数）特征、PLP(Perceptual Linear Predictive—感知线性预测)特征、或者 LPC（Linear Predictive Coding—线性预测编码）特征等。

下面以提取某一音频帧的 MFCC 特征为例，对特征矢量的提取过程作进一步描述。首先将音频帧的时域信号通过 FFT（Fast Fourier Transformation—快速傅氏变换）得到对应的频谱信息，将所述频谱信息通过 Mel 滤波器组得到 Mel 频谱，在 Mel 频谱上进行倒谱分析，其核心一般是采用 DCT（Discrete Cosine Transform—离散余弦变换）进行逆变换，然后取预设的 N 个系数（例如 N=12 或者 38），则得到了所述音频帧的特征矢量：

MFCC 特征。对每个音频帧都采用上述方式进行处理，可以得到表征所述语音信号的一系列特征矢量，即本申请所述的特征矢量序列。

步骤 103、计算特征矢量对应于搜索空间基本单元的概率。

在本实施例中，WFST 搜索空间基本单元是三音素，因此本步骤计算特征矢量对应于各三音素的概率。为了提高语音识别的准确率，本实施例采用 HMM 模型和具备强大特征提取能力的 DNN 模型计算所述概率，在其他实施方式中，也可以采用其他方式，例如：采用传统的 GMM 和 HMM 模型计算所述概率也同样可以实现本申请的技术方案，也在本申请的保护范围之内。

在具体实施时，可以在计算特征矢量对应于各三音素状态的基础上，进一步计算特征矢量对应于各三音素的概率，下面结合附图 6，对本步骤的处理过程作进一步描述。

步骤 103-1、采用预先训练的 DNN 模型计算特征矢量对应于各三音素状态的概率。

在本实施例的准备阶段已经预先训练好了 DNN 模型，本步骤以步骤 102 提取的特征矢量作为所述 DNN 模型的输入，则可以得到特征矢量对应于各三音素状态的概率。例如：三音素的数量为 1000，每个三音素包含 3 个状态，那么总共有 3000 个三音素状态，本步骤 DNN 模型输出：特征矢量对应于 3000 个三音素状态中每一状态的概率。

优选地，由于采用 DNN 模型涉及的计算量通常很大，本实施例通过利用硬件平台提供的数据并行处理能力提升采用 DNN 模型进行计算的速度。例如，目前嵌入式设备和移动设备用的比较多的是 ARM 架构平台，在现行的大多数 ARM 平台上，都有 SIMD（single instruction multiple data—单指令多数据流）的 NEON 指令集，该指令集可以在一个指令中处理多个数据，具备一定的数据并行处理能力，在本实施例中通过矢量化编程可以形成单指令流多数据流的编程泛型，从而可以充分利用硬件平台提供的数据并行处理能力，实现加速 DNN 计算的目的。

在客户端设备上实施本申请技术方案时，为了能够与客户端的硬件能力相匹配，通常会缩减 DNN 模型的规模，这样往往会导致 DNN 模型的精确度下降，对不同语音内容的识别能力也会随着下降。本实施例由于利用硬件加速机制，则可以不缩减或者尽量少缩减 DNN 模型的规模，从而可以最大限度地保留 DNN 模型的精确性，提高识别准确率。

步骤 103-2、根据特征矢量对应于所述各三音素状态的概率，采用预先训练的 HMM 模型计算特征矢量对应于各三音素的概率。

在准备阶段已经训练好了针对每个三音素的 HMM 模型，本步骤根据连续输入的若干个特征矢量对应于各三音素状态的概率，利用 HMM 模型计算对应于各三音素的转移

概率，从而得到特征矢量对应于各三音素的概率。

该计算过程实际上就是根据连续的特征矢量在各 HMM 上的传播过程计算相应转移概率的过程，下面以计算针对某一三音素（包括 3 个状态）的概率为例对计算过程作进一步说明，其中，pe(i,j)表示第 i 帧特征矢量在第 j 个状态上的发射概率，pt(h,k)表示从 h 状态转移到 k 状态的转移概率：

1）第一帧的特征矢量对应于相应 HMM 的状态 1，具有发射概率 pe(1,1)；

2）对于第二帧的特征矢量，如果从所述 HMM 的状态 1 转移到状态 1，对应的概率为 pe(1,1)*pt(1,1)*pe(2,1)，如果从状态 1 转移到状态 2，对应的概率为 pe(1,1)*pt(1,2)*pe(2,2)，根据上述概率判断转移到状态 1 还是状态 2；

3）对于第三帧的特征矢量以及后续帧的特征矢量也采用上述类似的计算方式，直到从状态 3 转移出去，至此在所述 HMM 上的传播结束，从而得到了连续多帧的特征矢量针对该 HMM 的概率，即：对应于该 HMM 表征的三音素的概率。

针对连续输入的特征矢量，采用上述方式计算在各 HMM 上传播的转移概率，从而得到对应于各三音素的概率。

步骤 104、以所述概率为输入、在所述搜索空间中执行解码操作，得到与所述特征矢量序列对应的词序列。

根据步骤 103 输出的特征矢量对应于各三音素的概率，在 WFST 网络中执行解码操作，得到与所述特征矢量序列对应的词序列。该过程通常为执行图搜索、找到得分最高的路径的搜索过程。常用的搜索方法是维特比算法，它的好处在于采用动态规划的方法节省了计算量，也可以做到时间同步解码。

考虑到在实际解码过程中，由于搜索空间巨大，维特比算法的计算量仍然很大，为了减少计算量、提高计算速度，在解码过程中并不是把所有路径的可能的后续路径都展开，而是只展开那些在最优路径附近的路径，即：可以在采用维特比算法进行搜索的过程中，采用适当的剪枝策略以提高搜索效率，例如：可以采用维特比柱搜索算法或者是采用直方图剪枝策略等。

至此，通过解码得到了与特征矢量序列对应的词序列，即，获取了待识别语音信号对应的识别结果。由于在步骤 101 构建用于进行语音识别的搜索空间时，添加了客户端预设信息，因此上述语音识别过程通常可以比较准确地识别出与客户端本地信息相关的语音内容。

考虑到客户端本地信息有可能被用户修改或者删除，为了进一步保证通过上述解码

过程获得的词序列的准确性，本实施例还提供一种优选实施方式：通过与所述客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

在具体实施时，上述优选实施方式可以包括以下所列的步骤 104-1 至步骤 104-4，下面结合附图 7 做进一步说明。

步骤 104-1、从所述词序列中选择对应于所述客户端预设信息的待验证词。

例如：针对打电话应用，所述客户端预设信息为通讯录中的联系人名称，语音识别的结果为词序列“给小明打电话”，那么通过与模板匹配的方式或者语法解析过程，可以确定所述词序列中的“小明”是与客户端预设信息对应的待验证词。

步骤 104-2、在所述客户端预设信息中查找所述待验证词，若找到，判定通过准确性验证，执行步骤 104-3，否则执行步骤 104-4。

本步骤通过执行文字层面的精准匹配，判断所述待验证词是否属于相对应的客户端预设信息，从而验证所述词序列的准确性。

仍沿用步骤 104-1 中的例子，本步骤在客户端通讯录中查找是否存在“小明”这个联系人，即：通讯录中与联系人名称相关的信息中是否包含“小明”这一字符串，若包含，则判定通过准确性验证，继续执行步骤 104-3，否则，转到步骤 104-4 执行。

步骤 104-3、将所述词序列作为语音识别结果。

执行到本步骤，说明通过解码得到的词序列中所包含的待验证词，与客户端预设信息是相吻合的，可以将所述词序列作为语音识别结果输出，从而触发使用该语音识别结果的应用程序执行相应的操作。

步骤 104-4、通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

执行到本步骤，通常说明所述通过解码得到的词序列中所包含的待验证词，与客户端预设信息是不相吻合的，如果将该词序列作为语音识别结果输出，那么相关应用程序通常无法执行正确的操作，因此在这种情况下，可以通过拼音层面的模糊匹配对所述词序列进行必要的修正。

在具体实施时，可以通过如下方式实现上述修正功能：通过查找发音词典，将所述待验证词转换为待验证拼音序列，将所述客户端预设信息中的各个词也分别转换为比对拼音序列，然后依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照相似度从高到低排序靠前的词，最后用所选词替换所述词

序列中的待验证词，得到所述修正后的词序列。

在具体实施时可以采用不同的方式计算两个拼音序列之间的相似度，本实施例采用基于编辑距离计算所述相似度的方式，例如：用两个拼音序列之间的编辑距离与 1 相加求和的倒数作为所述相似度。所述编辑距离（Edit Distance），是指两个字串之间，由一个转成另一个所需的最少编辑操作次数，所述编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符，一般来说，编辑距离越小，两个串的相似度越大。

仍沿用上述步骤 104-1 中的例子，词序列为“给小明打电话”，待验证词为“小明”，如果在客户端通讯录的联系人中没有找到“小明”，则可以通过查找发音词典，将小明转换为待验证拼音序列“xiaoming”，并将通讯录中的各个联系人名称也都转换为相应的拼音序列，即：比对拼音序列，然后依次计算“xiaoming”与各比对拼音序列之间的编辑距离，然后选择编辑距离最小（相似度最高）的比对拼音序列所对应的联系人名称（例如：“xiaomin”对应的“小敏”），替换所述词序列中的待验证词，从而完成了对所述词序列的修正，并可以将修正后的词序列作为最终的语音识别结果。

在具体实施时，也可以先计算出待验证拼音序列与各比对拼音序列之间的相似度并按照相似度从高到低排序，选择排序靠前的若干个（例如三个）比对拚音序列对应的词，然后将这些词通过屏幕输出等方式提示给客户端用户，由用户从中选择正确的词，并根据用户选择的词替换所述词序列中的待验证词。

至此，通过上述步骤 101-步骤 104 对本申请提供的语音识别方法的具体实施方式进行了详细说明。为了便于理解，请参考图 8，其为本实施例提供的语音识别过程的整体框架图。其中虚线框对应本实施例描述的准备阶段，实线框对应具体的语音识别处理过程。

需要说明的是，本实施例描述的步骤 101 可以在以语音作为交互媒介的客户端应用程序每次启动时都执行一次，即每次启动都重新生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间，也可以仅在所述客户端应用程序首次启动时生成所述搜索空间并存储、后续采用定期更新的方式，这样可以减少每次应用程序启动时生成搜索空间的时间开销（可以直接使用之前已生成的搜索空间），提高语音识别的执行效率，改善用户的使用体验。

此外，本申请提供的方法通常在客户端设备上实施，所述客户端设备包括：智能移动终端、智能音箱、机器人、或者其他可以运行所述方法的设备，本实施例即描述了在客户端实施本申请所提供方法的具体实施方式。但是在其他实施方式中，本申请提供的

方法也可以在基于客户端和服务器模式的应用场景下实施，在这种情况下，在准备阶段生成的各个 WFST 以及用于声学概率计算的模型无需预先安装到客户端设备中，每次客户端应用启动时，可以将相应的客户端预设信息上传给服务器，并将后续采集到待识别语音信号也上传给服务器，由服务器一侧实施本申请提供的方法，并将解码得到的词序列回传给客户端，同样可以实现本申请的技术方案，并取得相应的有益效果。

综上所述，本申请提供的语音识别方法，由于在生成用于对语音信号进行解码的搜索空间时包含了客户端预设信息，因此在对客户端采集的语音信号进行识别时能够相对准确地识别出与客户端本地相关的信息，从而可以提高语音识别的准确率，提升用户的使用体验。

特别是在客户端设备上采用本申请提供的方法进行语音识别，由于添加了客户端本地信息，因此可以在一定程度上弥补由于概率计算模型以及搜索空间规模缩小导致的识别准确率下降的问题，从而既能够满足在没有网络接入环境下进行语音识别的需求，同时也能达到一定的识别准确率。进一步地，在解码得到词序列后，通过采用本实施例给出的基于文字层面以及拼音层面的匹配验证方案，可以进一步提升语音识别的准确率。通过实际的测试结果表明：常规的语音识别方法的字符错误率（CER）在 20%左右，而使用本申请提供的方法，字符错误率在 3%以下，以上数据充分说明了本方法的有益效果是显著的。

在上述的实施例中，提供了一种语音识别方法，与之相对应的，本申请还提供一种语音识别装置。请参看图 9，其为本申请的一种语音识别装置的实施例的示意图。由于装置实施例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。下述描述的装置实施例仅仅是示意性的。

本实施例的一种语音识别装置，包括：搜索空间生成单元 901，用于利用预设的语音知识源，生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间；特征矢量提取单元 902，用于提取待识别语音信号的特征矢量序列；概率计算单元 903，用于计算特征矢量对应于搜索空间基本单元的概率；解码搜索单元 904，用于以所述概率为输入、在所述搜索空间中执行解码操作，得到与所述特征矢量序列对应的词序列。

可选的，所述搜索空间生成单元具体用于，通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息，并得到基于三音素状态绑定列表、发音词典、以及语言模型的单一加权有限状态转换器；

所述语言模型是由语言模型训练单元预先生成的，所述语言模型训练单元用于，将

用于训练语言模型的文本中的预设命名实体替换为与预设主题类别对应的标签，并利用所述文本训练语言模型。

可选的，所述搜索空间生成单元包括：

第一客户端信息添加子单元，用于通过替换标签的方式，向预先生成的基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息；

加权有限状态转换器合并子单元，用于将添加了所述客户端预设信息的所述加权有限状态转换器、与预先生成的基于三音素状态绑定列表和发音词典的加权有限状态转换器进行合并，得到所述单一加权有限状态转换器。

可选的，所述解码空间生成单元包括：

第二客户端信息添加子单元，用于通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息；

统一加权有限状态转换器获取子单元，用于在所述第二客户端信息添加子单元完成添加操作之后，得到基于三音素状态绑定列表、发音词典、以及语言模型的单一加权有限状态转换器；

其中，所述第二客户端信息添加子单元包括：

主题确定子单元，用于确定待识别语音信号所属的预设主题类别；

加权有限状态转换器选择子单元，用于选择预先生成的、与所述预设主题类别相对应的所述至少基于语言模型的加权有限状态转换器；

标签替换子单元，用于通过用与所述预设主题类别对应的客户端预设信息替换相应标签的方式，向所选的加权有限状态转换器中添加客户端预设信息。

可选的，所述主题确定子单元具体用于，根据采集所述语音信号的客户端类型、或应用程序确定所述所属的预设主题类别。

可选的，所述加权有限状态转换器合并子单元具体用于，采用基于预测的方法执行合并操作，并得到所述单一加权有限状态转换器。

可选的，所述概率计算单元包括：

三音素状态概率计算子单元，用于采用预先训练的 DNN 模型计算特征矢量对应于各三音素状态的概率；

三音素概率计算子单元，用于根据特征矢量对应于所述各三音素状态的概率，采用预先训练的 HMM 模型计算特征矢量对应于各三音素的概率。

可选的，所述特征矢量提取单元包括：

分帧子单元，用于按照预先设定的帧长度对待识别语音信号进行分帧处理，得到多个音频帧；

特征提取子单元，用于提取各音频帧的特征矢量，得到所述特征矢量序列。

可选的，所述装置包括：

准确性验证单元，用于在所述解码搜索单元得到与特征矢量序列对应的词序列后，通过与所述客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

可选的，所述准确性验证单元包括：

待验证词选择子单元，用于从所述词序列中选择对应于所述客户端预设信息的待验证词；

查找子单元，用于在所述客户端预设信息中查找所述待验证词；

识别结果确认子单元，用于当所述查找子单元找到所述待验证词之后，判定通过所述准确性验证，并将所述词序列作为语音识别结果；

识别结果修正子单元，用于当所述查找子单元未找到所述待验证词之后，通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

可选的，所述识别结果修正子单元，包括：

待验证拼音序列转换子单元，用于将所述待验证词转换为待验证拼音序列；

比对拼音序列转换子单元，用于将所述客户端预设信息中的各个词分别转换为比对拼音序列；

相似度计算选择子单元，用于依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；

待验证词替换子单元，用于用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

此外，本申请还提供另一种语音识别方法，请参考图 10，其为本申请提供的另一种语音识别方法的实施例的流程图，本实施例与之前提供的方法实施例内容相同的部分不再赘述，下面重点描述不同之处。本申请提供的另一种语音识别方法包括：

步骤 1001、通过解码获取与待识别语音信号对应的词序列。

对于语音识别来说，解码的过程就是在用于语音识别的搜索空间中进行搜索的过程，以获取与待识别语音信号对应的最佳词序列。所述搜索空间可以是基于各种知识源的

WFST 网络，也可以是其他形式的搜索空间；所述搜索空间可以包含客户端预设信息，也可以不包含客户端预设信息，本实施例并不对此作具体的限定。

步骤 1002、通过与客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

本步骤包括以下操作：从所述词序列中选择对应于所述客户端预设信息的待验证词；在所述客户端预设信息中查找所述待验证词；若找到，则判定通过所述准确性验证，并将所述词序列作为语音识别结果；否则通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

所述通过基于拼音的模糊匹配方式修正所述词序列，包括：将所述待验证词转换为待验证拼音序列；将所述客户端预设信息中的各个词分别转换为比对拼音序列；依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

其中，所述转换拼音序列可以通过查找发音词典实现，所述相似度可以根据两个拼音序列之间的编辑距离计算。

本申请提供的方法，通常应用于用语音作为交互媒介的应用程序中，此类应用程序采集的待识别语音中可能会涉及客户端信息，而本申请提供的方法，通过将解码得到的词序列与客户端预设信息进行文字匹配，可以验证所述词序列的准确性，从而为对词序列进行必要修正提供了依据。进一步地，通过采用基于拼音层面的模糊匹配，可以对所述词序列进行修正，从而提升语音识别的准确率。

在上述的实施例中，提供了另一种语音识别方法，与之相对应的，本申请还提供另一种语音识别装置。请参看图 11，其为本申请的另一种语音识别装置的实施例示意图。由于装置实施例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。下述描述的装置实施例仅仅是示意性的。

本实施例的一种语音识别装置，包括：词序列获取单元 1101，用于通过解码获取与待识别语音信号对应的词序列；词序列验证单元 1102，用于通过与客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

可选的，所述词序列验证单元包括：

待验证词选择子单元，用于从所述词序列中选择对应于所述客户端预设信息的待验证词；

查找子单元，用于在所述客户端预设信息中查找所述待验证词；

识别结果确认子单元，用于当所述查找子单元找到所述待验证词之后，判定通过所述准确性验证，并将所述词序列作为语音识别结果；

识别结果修正子单元，用于当所述查找子单元未找到所述待验证词之后，通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

可选的，所述识别结果修正子单元，包括：

待验证拼音序列转换子单元，用于将所述待验证词转换为待验证拼音序列；

比对拼音序列转换子单元，用于将所述客户端预设信息中的各个词分别转换为比对拼音序列；

相似度计算选择子单元，用于依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；

待验证词替换子单元，用于用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

本申请虽然以较佳实施例公开如上，但其并不是用来限定本申请，任何本领域技术人员在不脱离本申请的精神和范围内，都可以做出可能的变动和修改，因此本申请的保护范围应当以本申请权利要求所界定的范围为准。

在一个典型的配置中，计算设备包括一个或多个处理器 (CPU)、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器 (RAM) 和/或非易失性内存等形式，如只读存储器 (ROM) 或闪存(flash RAM)。内存是计算机可读介质的示例。

1、计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括非暂存电脑可读媒体

(transitory media)，如调制的数据信号和载波。

2、本领域技术人员应明白，本申请的实施例可提供为方法、系统或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

权 利 要 求 书

1、一种语音识别方法，其特征在于，包括：

利用预设的语音知识源，生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间；

提取待识别语音信号的特征矢量序列；

计算特征矢量对应于搜索空间基本单元的概率；

以所述概率为输入、在所述搜索空间中执行解码操作，得到与所述特征矢量序列对应的词序列。

2、根据权利要求 1 所述的语音识别方法，其特征在于，所述搜索空间包括：加权有限状态转换器。

3、根据权利要求 2 所述的语音识别方法，其特征在于，所述搜索空间基本单元包括：上下文相关的三音素；

所述预设的知识源包括：发音词典、语言模型、以及三音素状态绑定列表。

4、根据权利要求 3 所述的语音识别方法，其特征在于，所述利用预设的语音知识源生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间，包括：

通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息，并得到基于三音素状态绑定列表、发音词典以及语言模型的单一加权有限状态转换器；

其中，所述语言模型是采用如下方式预先训练得到的：

将用于训练语言模型的文本中的预设命名实体替换为与预设主题类别对应的标签，并利用所述文本训练语言模型。

5、根据权利要求 4 所述的语音识别方法，其特征在于，所述通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息，并得到基于三音素状态绑定列表、发音词典以及语言模型的单一加权有限状态转换器，包括：

通过替换标签的方式，向预先生成的基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息；

将添加了客户端预设信息的所述加权有限状态转换器、与预先生成的基于三音素状态绑定列表和发音词典的加权有限状态转换器进行合并，得到所述单一加权有限状态转换器。

6、根据权利要求 4 所述的语音识别方法，其特征在于，所述用于训练语言模型的文本是指，针对所述预设主题类别的文本。

7、根据权利要求 4 所述的语音识别方法，其特征在于，所述预设主题类别的数目至少为 2 个；所述语言模型的数目、以及所述至少基于语言模型的加权有限状态器的数目分别与预设主题类别的数目一致；

所述通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息，包括：

确定待识别语音信号所属的预设主题类别；

选择预先生成的、与所述预设主题类别相对应的所述至少基于语言模型的加权有限状态转换器；

通过用与所述预设主题类别对应的客户端预设信息替换相应标签的方式，向所选的加权有限状态转换器中添加客户端预设信息。

8、根据权利要求 7 所述的语音识别方法，其特征在于，所述确定待识别语音信号所属的预设主题类别，采用如下方式实现：

根据采集所述语音信号的客户端类型、或应用程序确定所述所属的预设主题类别。

9、根据权利要求 8 所述的语音识别方法，其特征在于，所述预设主题类别包括：拨打电话或发送短信，播放乐曲，或者，设置指令；

相应的客户端预设信息包括：通讯录中的联系人名称，曲库中的乐曲名称，或者，指令集中的指令。

10、根据权利要求 5 所述的语音识别方法，其特征在于，所述合并操作包括：采用基于预测的方法进行合并。

11、根据权利要求 4 所述的语音识别方法，其特征在于，预先训练所述语言模型所采用的词表与所述发音词典包含的词一致。

12、根据权利要求 3 所述的语音识别方法，其特征在于，所述计算特征矢量对应于搜索空间基本单元的概率，包括：

采用预先训练的 DNN 模型计算特征矢量对应于各三音素状态的概率；

根据特征矢量对应于所述各三音素状态的概率，采用预先训练的 HMM 模型计算特征矢量对应于各三音素的概率。

13、根据权利要求 12 所述的语音识别方法，其特征在于，通过如下方式提升所述采用预先训练的 DNN 模型计算特征矢量对应于各三音素状态的概率的步骤的执行速

度：利用硬件平台提供的数据并行处理能力。

14、根据权利要求 1-13 任一项所述的语音识别方法，其特征在于，所述提取待识别语音信号的特征矢量序列，包括：

按照预先设定的帧长度对待识别语音信号进行分帧处理，得到多个音频帧；

提取各音频帧的特征矢量，得到所述特征矢量序列。

15、根据权利要求 14 所述的语音识别方法，其特征在于，所述提取各音频帧的特征矢量包括：提取 MFCC 特征、PLP 特征、或者 LPC 特征。

16、根据权利要求 1-13 任一项所述的语音识别方法，其特征在于，在所述得到与所述特征矢量序列对应的词序列后，执行下述操作：

通过与所述客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

17、根据权利要求 16 所述的语音识别方法，其特征在于，所述通过与所述客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果得到相应的语音识别结果，包括：

从所述词序列中选择对应于所述客户端预设信息的待验证词；

在所述客户端预设信息中查找所述待验证词；

若找到，则判定通过所述准确性验证，并将所述词序列作为语音识别结果；否则通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

18、根据权利要求 17 所述的语音识别方法，其特征在于，所述通过基于拼音的模糊匹配方式修正所述词序列，包括：

将所述待验证词转换为待验证拼音序列；

将所述客户端预设信息中的各个词分别转换为比对拼音序列；

依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；

用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

19、根据权利要求 18 所述的语音识别方法，其特征在于，所述相似度包括：基于编辑距离计算的相似度。

20、根据权利要求 1-13 任一项所述的语音识别方法，其特征在于，所述方法在客户端设备上实施；所述客户端设备包括：智能移动终端、智能音箱、或者机器人。

21、一种语音识别装置，其特征在于，包括：

搜索空间生成单元，用于利用预设的语音知识源，生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间；

特征矢量提取单元，用于提取待识别语音信号的特征矢量序列；

概率计算单元，用于计算特征矢量对应于搜索空间基本单元的概率；

解码搜索单元，用于以所述概率为输入、在所述搜索空间中执行解码操作，得到与所述特征矢量序列对应的词序列。

22、根据权利要求 21 所述的语音识别装置，其特征在于，所述搜索空间生成单元具体用于，通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息，并得到基于三音素状态绑定列表、发音词典、以及语言模型的单一加权有限状态转换器；

所述语言模型是由语言模型训练单元预先生成的，所述语言模型训练单元用于，将用于训练语言模型的文本中的预设命名实体替换为与预设主题类别对应的标签，并利用所述文本训练语言模型。

23、根据权利要求 22 所述的语音识别装置，其特征在于，所述搜索空间生成单元包括：

第一客户端信息添加子单元，用于通过替换标签的方式，向预先生成的基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息；

加权有限状态转换器合并子单元，用于将添加了所述客户端预设信息的所述加权有限状态转换器、与预先生成的基于三音素状态绑定列表和发音词典的加权有限状态转换器进行合并，得到所述单一加权有限状态转换器。

24、根据权利要求 22 所述的语音识别装置，其特征在于，所述解码空间生成单元包括：

第二客户端信息添加子单元，用于通过替换标签的方式，向预先生成的至少基于语言模型的加权有限状态转换器中添加与预设主题类别对应的客户端预设信息；

统一加权有限状态转换器获取子单元，用于在所述第二客户端信息添加子单元完成添加操作之后，得到基于三音素状态绑定列表、发音词典、以及语言模型的单一加权有限状态转换器；

其中，所述第二客户端信息添加子单元包括：

主题确定子单元，用于确定待识别语音信号所属的预设主题类别；

加权有限状态转换器选择子单元，用于选择预先生成的、与所述预设主题类别相对

应的所述至少基于语言模型的加权有限状态转换器；

标签替换子单元，用于通过用与所述预设主题类别对应的客户端预设信息替换相应标签的方式，向所选的加权有限状态转换器中添加客户端预设信息。

25、根据权利要求 24 所述的语音识别装置，其特征在于，所述主题确定子单元具体用于，根据采集所述语音信号的客户端类型、或应用程序确定所述所属的预设主题类别。

26、根据权利要求 23 所述的语音识别装置，其特征在于，所述加权有限状态转换器合并子单元具体用于，采用基于预测的方法执行合并操作，并得到所述单一加权有限状态转换器。

27、根据权利要求 21 所述的语音识别装置，其特征在于，所述概率计算单元包括：

三音素状态概率计算子单元，用于采用预先训练的 DNN 模型计算特征矢量对应于各三音素状态的概率；

三音素概率计算子单元，用于根据特征矢量对应于所述各三音素状态的概率，采用预先训练的 HMM 模型计算特征矢量对应于各三音素的概率。

28、根据权利要求 21-27 任一项所述的语音识别装置，其特征在于，所述特征矢量提取单元包括：

分帧子单元，用于按照预先设定的帧长度对待识别语音信号进行分帧处理，得到多个音频帧；

特征提取子单元，用于提取各音频帧的特征矢量，得到所述特征矢量序列。

29、根据权利要求 21-27 任一项所述的语音识别装置，其特征在于，包括：

准确性验证单元，用于在所述解码搜索单元得到与特征矢量序列对应的词序列后，通过与所述客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

30、根据权利要求 29 所述的语音识别装置，其特征在于，所述准确性验证单元包括：

待验证词选择子单元，用于从所述词序列中选择对应于所述客户端预设信息的待验证词；

查找子单元，用于在所述客户端预设信息中查找所述待验证词；

识别结果确认子单元，用于当所述查找子单元找到所述待验证词之后，判定通过所述准确性验证，并将所述词序列作为语音识别结果；

识别结果修正子单元，用于当所述查找子单元未找到所述待验证词之后，通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

31、根据权利要求 30 所述的语音识别装置，其特征在于，所述识别结果修正子单元，包括：

待验证拼音序列转换子单元，用于将所述待验证词转换为待验证拼音序列；

比对拼音序列转换子单元，用于将所述客户端预设信息中的各个词分别转换为比对拼音序列；

相似度计算选择子单元，用于依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；

待验证词替换子单元，用于用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

32、一种语音识别方法，其特征在于，包括：

通过解码获取与待识别语音信号对应的词序列；

通过与客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

33、根据权利要求 32 所述的语音识别方法，其特征在于，所述通过与客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果，包括：

从所述词序列中选择对应于所述客户端预设信息的待验证词；

在所述客户端预设信息中查找所述待验证词；

若找到，则判定通过所述准确性验证，并将所述词序列作为语音识别结果；否则通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

34、根据权利要求 33 所述的语音识别方法，其特征在于，所述通过基于拼音的模糊匹配方式修正所述词序列，包括：

将所述待验证词转换为待验证拼音序列；

将所述客户端预设信息中的各个词分别转换为比对拼音序列；

依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；

用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

35、一种语音识别装置，其特征在于，包括：

词序列获取单元，用于通过解码获取与待识别语音信号对应的词序列；

词序列验证单元，用于通过与客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果。

36、根据权利要求 35 所述的语音识别装置，其特征在于，所述词序列验证单元包括：

待验证词选择子单元，用于从所述词序列中选择对应于所述客户端预设信息的待验证词；

查找子单元，用于在所述客户端预设信息中查找所述待验证词；

识别结果确认子单元，用于当所述查找子单元找到所述待验证词之后，判定通过所述准确性验证，并将所述词序列作为语音识别结果；

识别结果修正子单元，用于当所述查找子单元未找到所述待验证词之后，通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果。

37、根据权利要求 36 所述的语音识别装置，其特征在于，所述识别结果修正子单元，包括：

待验证拼音序列转换子单元，用于将所述待验证词转换为待验证拼音序列；

比对拼音序列转换子单元，用于将所述客户端预设信息中的各个词分别转换为比对拼音序列；

相似度计算选择子单元，用于依次计算所述待验证拼音序列与各比对拼音序列之间的相似度，并从所述客户端预设信息中选择按照所述相似度从高到低排序靠前的词；

待验证词替换子单元，用于用所选词替换所述词序列中的待验证词，得到所述修正后的词序列。

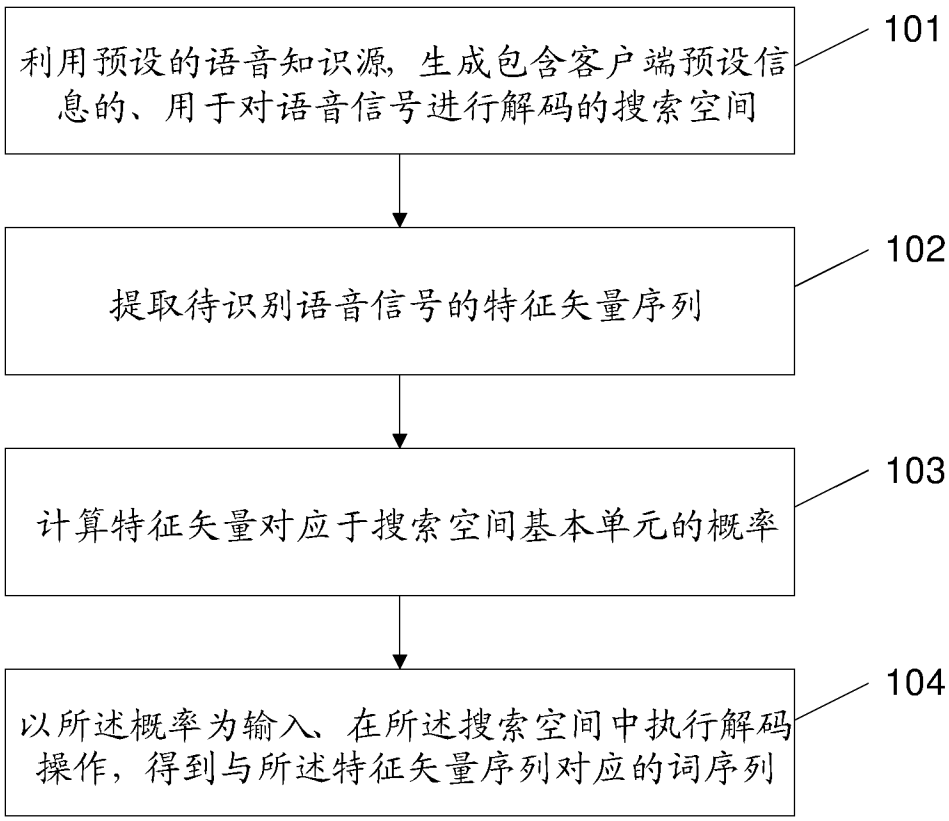
利用预设的语音知识源，生成包含客户端预设信息的、用于对语音信号进行解码的搜索空间
101
提取待识别语音信号的特征矢量序列
102
计算特征矢量对应于搜索空间基本单元的概率
103
以所述概率为输入、在所述搜索空间中执行解码操作，得到与所述特征矢量序列对应的词序列
104

图 1

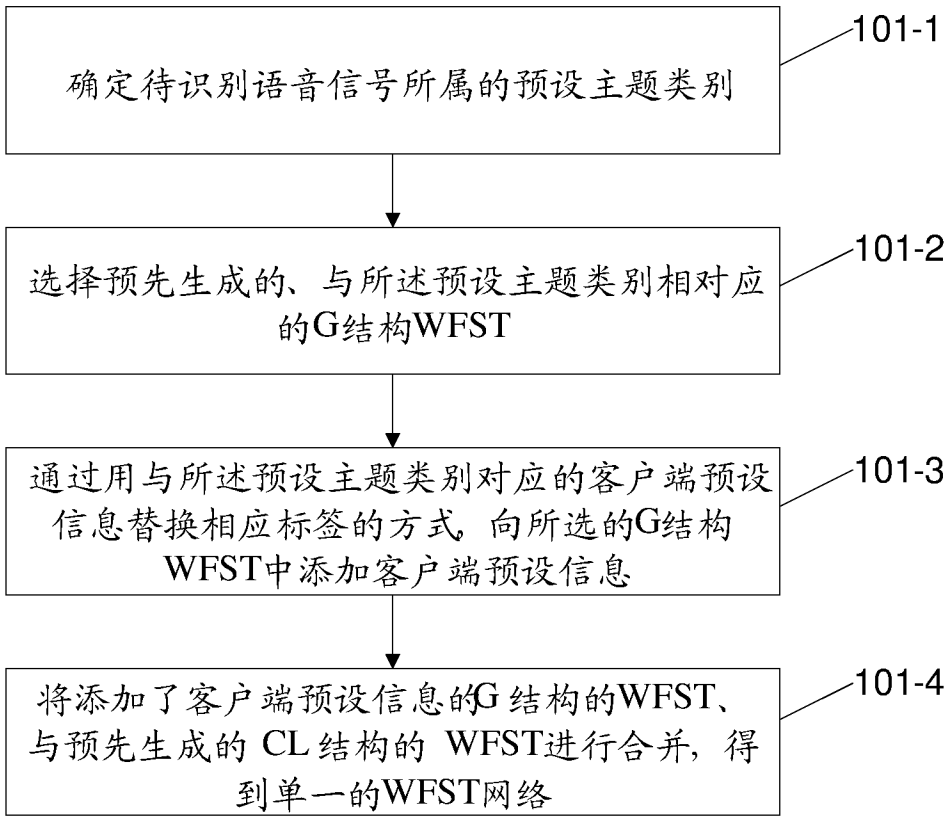
确定待识别语音信号所属的预设主题类别
101-1
选择预先生成的、与所述预设主题类别相对应的G结构WFST
101-2
通过用与所述预设主题类别对应的客户端预设信息替换相应标签的方式，向所选的G结构WFST中添加客户端预设信息
101-3
将添加了客户端预设信息的G结构的WFST、与预先生成的 CL 结构的 WFST进行合并，得到单一的WFST网络
101-4

图 2

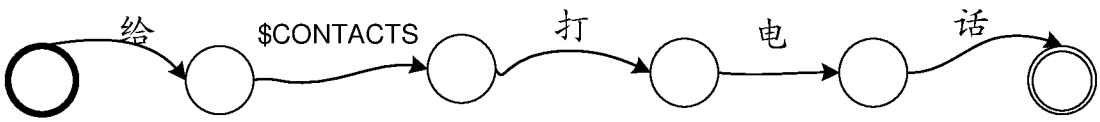
给
$CONTACTS
打
电
话

图 3

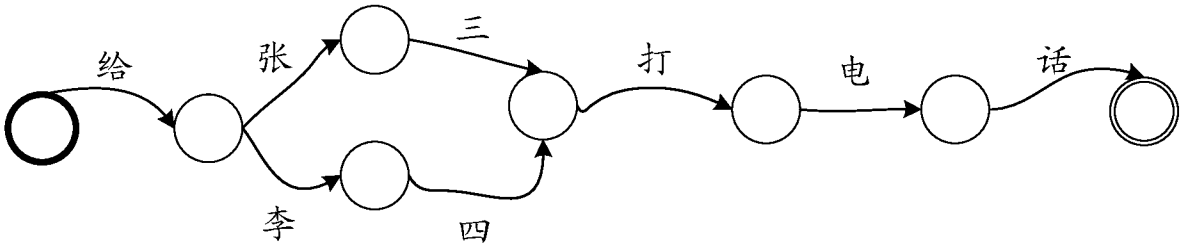
给
张
三
打
电
话
李
四

图 4

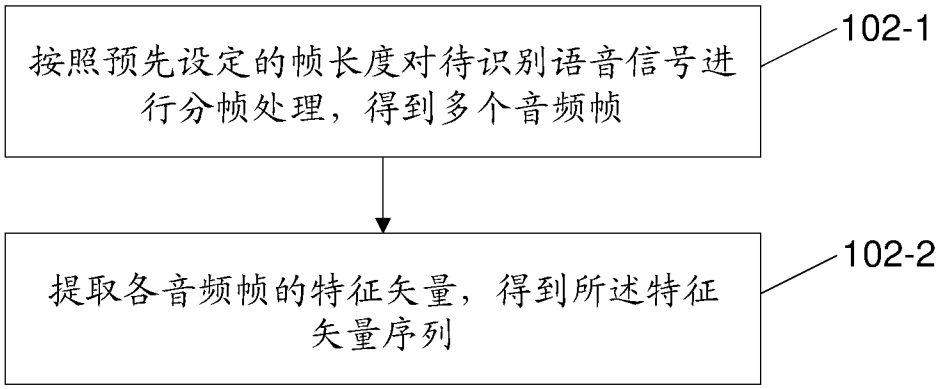
按照预先设定的帧长度对待识别语音信号进行分帧处理，得到多个音频帧
102-1
提取各音频帧的特征矢量，得到所述特征矢量序列
102-2

图 5

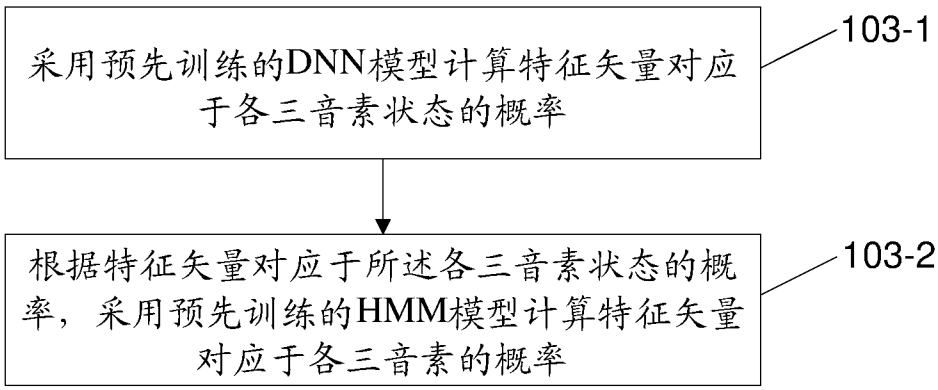
采用预先训练的DNN模型计算特征矢量对应于各三音素状态的概率
103-1
根据特征矢量对应于所述各三音素状态的概率，采用预先训练的HMM模型计算特征矢量对应于各三音素的概率
103-2

图 6

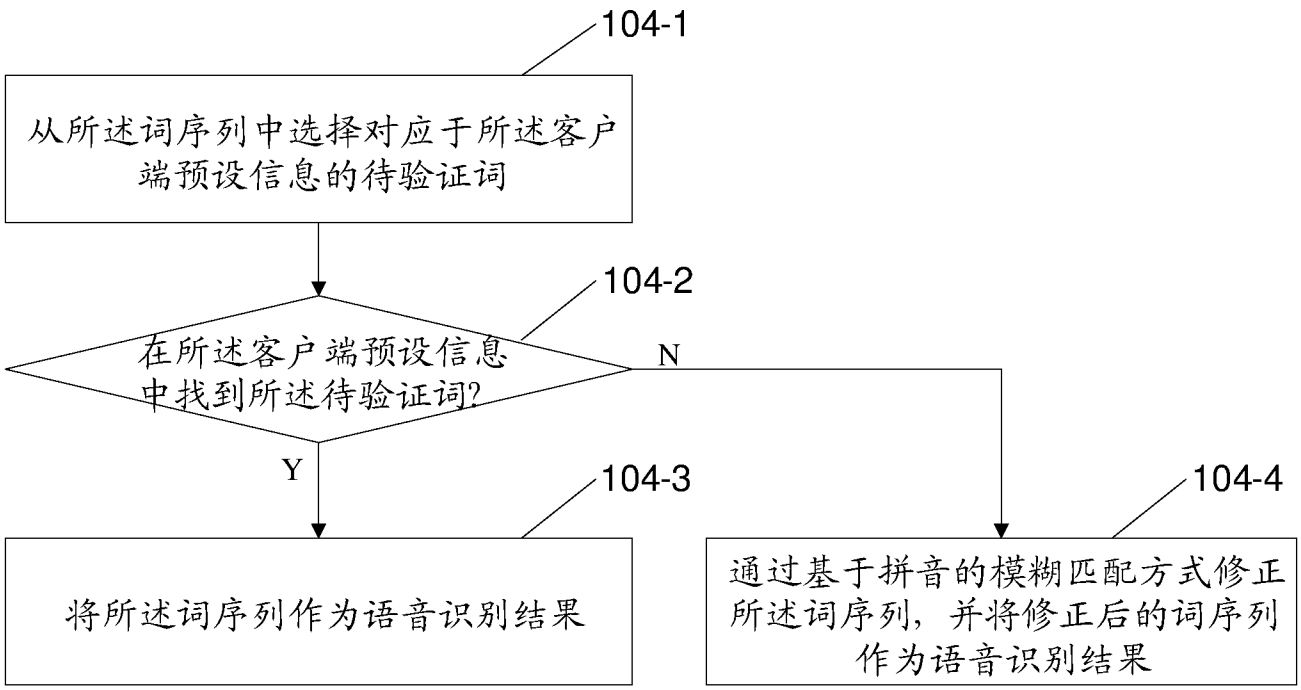
104-1
从所述词序列中选择对应于所述客户端预设信息的待验证词
104-2
在所述客户端预设信息中找到所述待验证词?
N
Y
104-3
将所述词序列作为语音识别结果
104-4
通过基于拼音的模糊匹配方式修正所述词序列，并将修正后的词序列作为语音识别结果

图 7

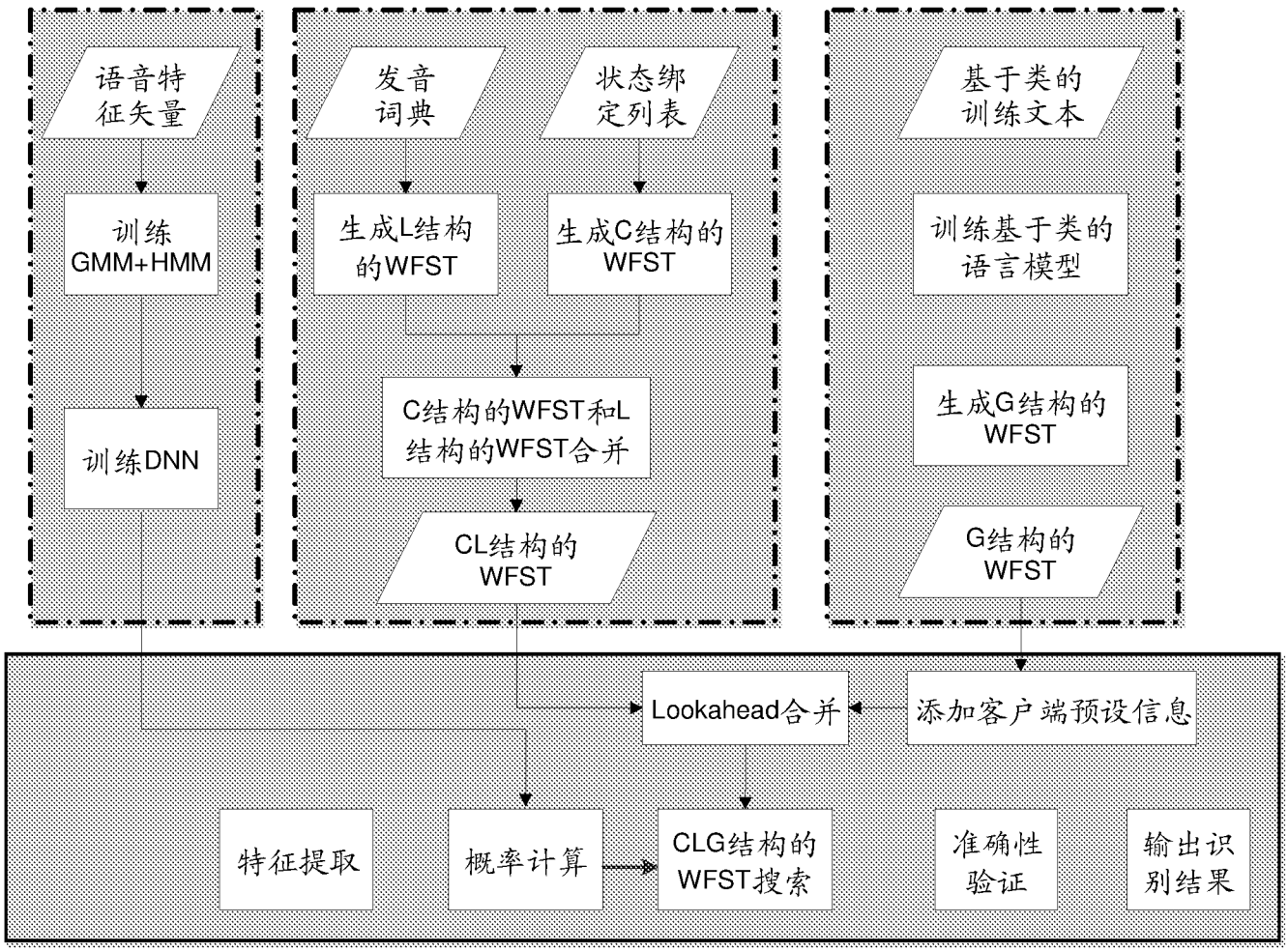
语音特征矢量
训练GMM+HMM
训练DNN
发音词典
状态绑定列表
生成L结构的WFST
生成C结构的WFST
C结构的WFST和L结构的WFST合并
CL结构的WFST
基于类的训练文本
训练基于类的语言模型
生成G结构的WFST
G结构的WFST
Lookahead合并
添加客户端预设信息
特征提取
概率计算
CLG结构的WFST搜索
准确性验证
输出识别结果

图 8

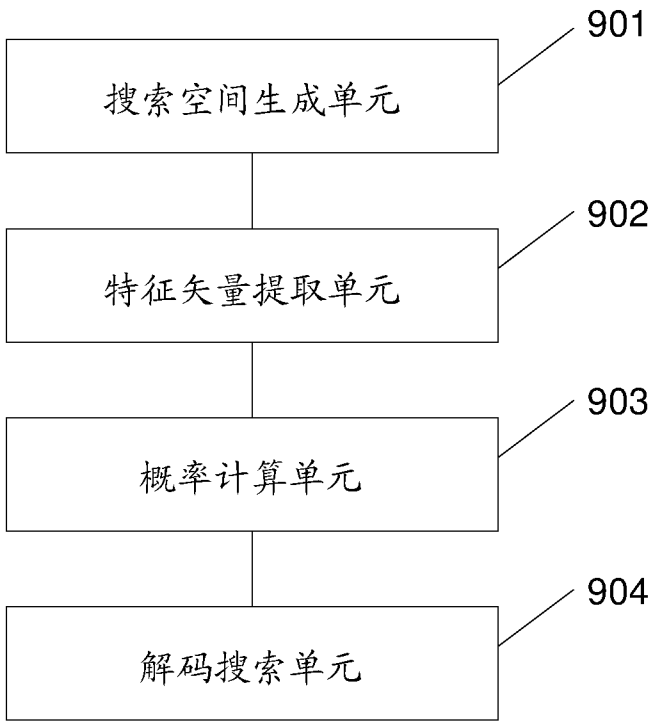
901
搜索空间生成单元
902
特征矢量提取单元
903
概率计算单元
904
解码搜索单元

图 9

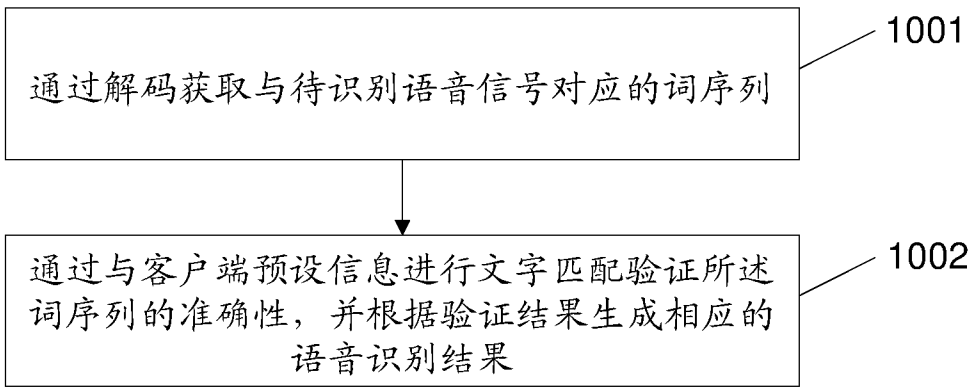
1001
通过解码获取与待识别语音信号对应的词序列
1002
通过与客户端预设信息进行文字匹配验证所述词序列的准确性，并根据验证结果生成相应的语音识别结果

图 10

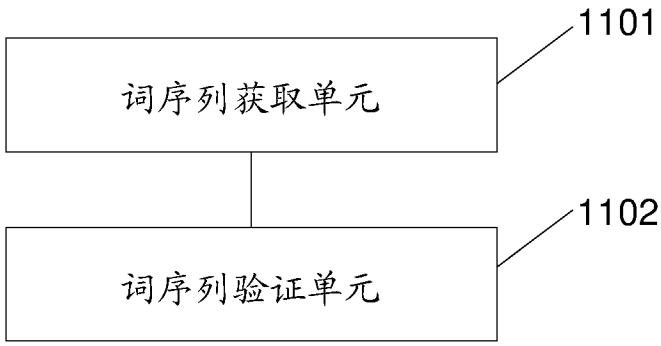
1101
词序列获取单元
1102
词序列验证单元

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016103691

A. CLASSIFICATION OF SUBJECT MATTER

G10L 15/30 (2013.01) i; G06F 17/27 (2006.01) 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 15; G06F 17

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)

CNABS, CNKI, CNTXT, VEN: personality, individual, reserve, pre-set, user, situation, search the Internet, searching space, accuracy, audio, speech, speak+, voice, sound, acous+, audible, recogn+, identify+, decod+, customize+, specif+, special+, adapt+, client?, terminal?, device?, application?, scene?, net?, network?, space, model

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KR 20150041599 A (SAMSUNG ELECTRONICS CO., LTD.), 16 April 2015 (16.04.2015), description, paragraphs [0020]-[0052]	1-37
X	US 2002138274 A1 (SHARMA, S.R. et al.), 26 September 2002 (26.09.2002), description, paragraphs [0032]-[0033] and [0040]-[0042]	1-37
X	US 8805684 B1 (GOOGLE INC.), 12 August 2014 (12.08.2014), description, column 13, line 65 to column 19, line 41	1-37
A	CN 103903619 A (ANHUI USTC IFLYTEK CO., LTD.), 02 July 2014 (02.07.2014), the whole document	1-37
A	US 2009125307 A1 (WANG, J.C.), 14 May 2009 (14.05.2009), the whole document	1-37

☒ 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
23 January 2017 (23.01.2017)	**04 February 2017 (04.02.2017)**
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer **YANG, Shilin** Telephone No.: (86-10) **62085717**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016103691

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 104538031 A (BEIJING UNISOUND INFORMATION TECHNOLOGY CO., LTD.), 22 April 2015 (22.04.2015), the whole document	1-37
A	CN 101320561 A (CYBERON CORP.), 10 December 2008 (10.12.2008), the whole document	1-37

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016103691

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
KR 20150041599 A	16 April 2015	US 2016232894 A1	11 August 2016
		CN 105814628 A	27 July 2016
		WO 2015053560 A1	16 April 2015
		IN 201627011312 A	15 July 2016
US 2002138274 A1	26 September 2002	None	
US 8805684 B1	12 August 2014	None	
CN 103903619 A	02 July 2014	WO 2014101826 A1	03 July 2014
US 2009125307 A1	14 May 2009	JP 2009145435 A	02 July 2009
CN 104538031 A	22 April 2015	None	
CN 101320561 A	10 December 2008	None	

国际检索报告

国际申请号
PCT/CN2016/103691

A. 主题的分类

G10L 15/30(2013.01)i; G06F 17/27(2006.01)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 15; G06F 17

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））

CNABS，CNKI，CNTXT，VEN：语音，声音，话音，识别，辨识，专用，定制，特性，特有，个性，个体，预定，预设，客户端，终端，用户，设备，装置，场景，场合，情景，情境，应用，搜索网络，搜索空间，模型，准确，精确，audio, speech, speak+, voice, sound, acous+, audible, recogn+, identify+, decod+, customize+, specif+, special+, adapt+, client?, terminal?, device?, application?, scene?, net?, network?, space, 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
X	KR 20150041599 A (SAMSUNG ELECTRONICS CO., LTD.) 2015年 4月 16日 (2015-04-16) 说明书第[0020]-[0052]段	1-37
X	US 2002138274 A1 (SHARMA S. R. 等) 2002年 9月 26日 (2002-09-26) 说明书第[0032]-[0033]段，第[0040]-[0042]段	1-37
X	US 8805684 B1 (GOOGLE INC.) 2014年 8月 12日 (2014-08-12) 说明书第13栏第65行-第19栏第41行	1-37
A	CN 103903619 A (安徽科大讯飞信息科技股份有限公司) 2014年 7月 2日 (2014-07-02) 全文	1-37
A	US 2009125307 A1 (WANG JUI-CHANG) 2009年 5月 14日 (2009-05-14) 全文	1-37

☑ 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
2017年 1月 23日	2017年 2月 4日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(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 杨士林 电话号码 (86-10)62085717

国际检索报告

国际申请号
PCT/CN2016/103691

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
A	CN 104538031 A (北京云知声信息技术有限公司) 2015年 4月 22日 (2015 - 04 - 22) 全文	1-37
A	CN 101320561 A (赛微科技股份有限公司) 2008年 12月 10日 (2008 - 12 - 10) 全文	1-37

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/103691

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)
KR	20150041599	A	2015年 4月 16日	US	2016232894	A1	2016年 8月 11日
				CN	105814628	A	2016年 7月 27日
				WO	2015053560	A1	2015年 4月 16日
				IN	201627011312	A	2016年 7月 15日
US	2002138274	A1	2002年 9月 26日		无		
US	8805684	B1	2014年 8月 12日		无		
CN	103903619	A	2014年 7月 2日	WO	2014101826	A1	2014年 7月 3日
US	2009125307	A1	2009年 5月 14日	JP	2009145435	A	2009年 7月 2日
CN	104538031	A	2015年 4月 22日		无		
CN	101320561	A	2008年 12月 10日		无		