(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 102324233 B
(45)授权公告日 2014.05.07

(21)申请号 201110220842.4
(22)申请日 2011.08.03
(73)专利权人 中国科学院计算技术研究所
地址 100190 北京市海淀区中关村科学院南路 6 号
(72)发明人 李新辉 王向东 钱跃良 林守勋
(74)专利代理机构 北京泛华伟业知识产权代理有限公司 11280
代理人 王勇
(51)Int.Cl.
G10L 15/26(2006.01)
G06F 17/30(2006.01)

(56)对比文件
JP 特开 2008-51895 A,2008.03.06,
CN 101923854 A,2010.12.22,
JP 特开 2005-227510 A,2005.08.25,
CN 101669116 A,2010.03.10,
US 2009/0313016 A1,2009.12.17,

审查员 陈红红

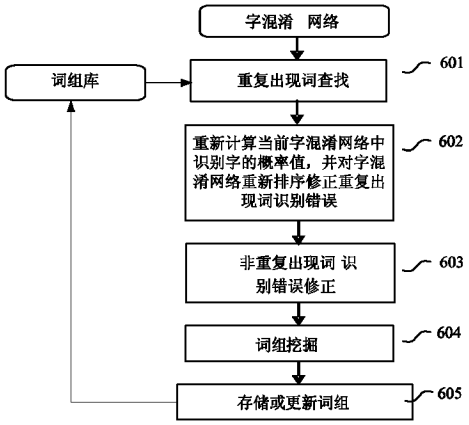
权利要求书2页 说明书9页 附图4页

(54)发明名称

汉语语音识别中重复出现词识别错误的自动修正方法

(57)摘要

本发明提供一种汉语语音识别中重复出现词识别错误的自动修正方法，包括：(1) 对每句话经识别后得到的字混淆网络与词组库中的词组及中间识别结果进行相似性匹配，以查找重复出现词组；其中，字混淆网络是所有可能识别结果的集合，字混淆网络包括最优识别结果即原有最优识别结果和最优识别结果中的每个字对应的中间识别结果；词组库包括词组及其对应的中间识别结果；(2) 根据查找得到的词组信息，重新计算相似概率值和字识别概率值；(3) 根据新的概率值，对字混淆网络按照概率值大小排序；和 (4) 使用排序结果替换字混淆网络的最优识别结果以及中间识别结果。优点在于：利用之前已修正的识别结果中的经验知识，自动修正当前识别语句中重复出现词的识别错误，从而提高识别错误的修正效率，加快识别错误修正速度。



CN 102324233 B

1. 一种汉语语音识别中重复出现词识别错误的自动修正方法，其特征在于，包括：

（1）对每句话经识别后得到的字混淆网络与词组库中的词组及中间识别结果进行相似性匹配，以查找重复出现词组；其中，字混淆网络是所有可能识别结果的集合，字混淆网络包括最优识别结果即原有最优识别结果和最优识别结果中的每个字对应的中间识别结果；所述词组库用于存储已修正识别结果中的词组及其对应的中间识别结果；所述已修正识别结果包括正确识别结果及用户修改识别结果；

（2）根据查找得到的词组信息，重新计算查找到的词组中每个字的相似概率值以及重新计算词组中的每个字所对应的当前字混淆网络中的一列识别结果中的字识别概率值；其中，所述词组信息包括词组本身、词组的相似概率值以及词组在最优识别结果中的对应位置；

（3）根据新的概率值，对字混淆网络按照概率值大小排序；

（4）使用排序结果替换字混淆网络的最优识别结果以及中间识别结果；

（5）修正非重复出现词识别错误，以得到不再包含任何识别错误的已修正识别结果；

（6）挖掘已修正识别结果中所有的词组；和

（7）存储或更新得到的词组到词组库。

2. 根据权利要求 1 所述的自动修正方法，其特征在于，步骤（1）中所述进行相似性匹配包括：

计算词组及对应中间识别结果与当前字混淆网络的相似概率值；和

保留相似概率值大于零的词组；

其中，所述计算方式为：

$$p(W,CN_j)=\prod_{i=1}^{num(W)}SIM(S_i,S'_{i+j})$$

$$SIM(C,C')=\frac{1}{2}\left[\frac{1}{N_1}\sum_{i=0}^{N_1-1}\delta(c_i,C')+\frac{1}{N_2}\sum_{i=0}^{N_2-1}\delta(c'_i,C)\right]$$

num(W) 表示词组 W 中字的个数，$p(W,CN_j)$ 表示词组对应中间识别结果与当前字混淆网络中第 j 列开始的 num(W) 列中间识别结果的相似概率，$SIM(S_i,S_{i+j}')$ 表示词组中第 i 个字所对应的中间识别结果与当前字混淆网络中第 i+j 列中间识别结果的相似性；

C 和 C′ 分别表示一列中间识别结果集合，N_1 和 N_2 分别表示 C 和 C′ 中字的个数；c_i 表示 C 中的第 i 个字，c_i' 表示 C′ 的第 i 个字；$\delta(c_i,C')$ 表示如果在 C′ 集合中存在某个字的读音与字 c_i 的读音相同，其值等于 1，否则等于 0。

3. 根据权利要求 1 所述的自动修正方法，其特征在于，步骤（2）中

重新计算查找到词组中每个字的相似概率值方式为：

$$p_{c_i}=\begin{cases}\lambda p'_{c_i}+(1-\lambda)q'(c_{loc+i,k}) & c_{loc+i,k}=c_i\\ \lambda p'_{c_i} & else\end{cases},$$

其中，p_{ci}' 表示该词组中第 i 个字的概率值，$q'(c_{loc+i,k})$ 表示第 i 个字所对应的第 loc+i 列识别结果中第 k 个字的原有识别概率值，$c_{loc+i,k}=c_i$ 表示第 i 个字所对应的第 loc+i 列识别结果中存在一个与之相同的字，λ 表示该词组为重复出现词的权重。

4. 根据权利要求 3 所述的自动修正方法，其特征在于，步骤（2）中

重新计算每个对应列中的字识别概率值的方式为：

$$q(c_{loc+i,j}) = (1-\lambda)q'(c_{loc+i,j}) + \lambda(1-p'_{c_i})q'(c_{loc+i,j})$$

p_{ci}' 表示词组 w 中第 i 个字的概率值，p_{ci}' 等于 p_w，p_w 表示词组 w 与当前字混淆网络的相似概率，$q'(c_{loc+i,j})$ 表示第 i 个字所对应的第 loc+i 列识别结果中第 j 个字的原有识别概率值，相应的 $q(c_{loc+i,j})$ 表示重新计算后的识别概率值，λ 表示词组 w 为重复出现词的权重。

5. 根据权利要求 1 所述的自动修正方法，其特征在于，所述步骤（5）通过在混淆网络中选择正确的字、或者通过键盘输入、或者通过手写输入来修正非重复出现词识别错误。

6. 根据权利要求 1 所述的自动修正方法，其特征在于，步骤（6）中所述挖掘已修正识别结果中所有的词组包括：

计算已修正识别结果中每个字与相邻若干字组成词的概率值；

选择组合概率最大的词组作为挖掘到的词组；

其中，所述计算的方式为：

$$p(c_1,c_2,\cdots,c_k) = \sum_{i_1=1}^{num(S_{c_1})} \sum_{i_2=1}^{num(S_{c_2})} \cdots \sum_{i_k=1}^{num(S_{c_k})} p(c_{1,i_1},c_{2,i_2},\cdots,c_{k,i_k})$$

$p(c_1, c_2, \ldots, c_k)$ 表示已修正识别结果中第 1 个字与第 2 个到第 k 个字组成词的概率值，$num(S_{ck})$ 表示已修正识别结果中第 k 个字对应中间识别结果列中字的个数，$c_{k,ik}$ 表示已修正识别结果中第 k 个字所对应中间识别结果列中的第 i_k 个字，$p(c_{1,i1}, c_{2,i2}, \cdots, c_{k,ik})$ 表示字混淆网络中字 $c_{1,i1}$ 与 $c_{2,i2}$ 到 $c_{k,ik}$ 的组合概率值。

7. 根据权利要求 1 所述的自动修正方法，其特征在于，所述词组库为词组文件或词组数据库。

汉语语音识别中重复出现词识别错误的自动修正方法

技术领域

[0001] 本发明涉及语音识别技术领域，特别是涉及一种汉语语音识别中重复出现词识别错误的自动修正方法。

背景技术

[0002] 语音识别技术是一种利用计算机和数字信号处理技术准确地识别出人类语音内容的技术。目前面向特殊应用的中小词汇量语音识别技术已得到实际应用，然而，由于受到背景噪音、方言口音、口语化的自然语音以及语义理解等因素的限制，大词汇量说话人无关的连续语音识别技术还处在探索阶段。由于语音识别无法达到 100％的识别准确率，因此，对识别结果中的识别错误进行修正是不可缺少的。

[0003] 识别错误修正是指在一句话识别后由说话人对识别结果中的错误进行修正。早期的识别错误修正方法主要有重新发音修正方法 (re-speaking)，单词拼写修正方法 (spelling)，键盘输入修正方法，和手写输入修正方法。近期的识别错误修正方法有候选选择修正方法，识别系统对每个词给出多个候选，用户在语音输入的同时或完成之后通过选择候选修正识别错误。无论是早期的修正方法还是后来的候选选择修正方法，对于不同语句中的同一个词识别错误都需要重新修正，即对于同一个词，其每次的识别错误都需要有用户参与的修正 ；修正效率较低。

发明内容

[0004] 本发明要解决的技术问题是利用之前已修正的识别结果，自动修正当前识别语句中重复出现词的识别错误，从而提高识别错误的修正效率，加快识别错误修正速度。

[0005] 本发明提供一种汉语语音识别中重复出现词识别错误的自动修正方法，其特征在于，包括 ：(1) 对每句话经识别后得到的字混淆网络与词组库中的词组及中间识别结果进行相似性匹配，以查找重复出现词组 ；(2) 根据查找得到的词组信息，重新计算相似概率值和字识别概率值 ；(3) 根据新的概率值，对字混淆网络按照概率值大小排序 ；和 (4) 使用排序结果替换字混淆网络的最优识别结果以及中间识别结果。

[0006] 其中，字混淆网络是所有可能识别结果的集合，字混淆网络包括最优识别结果即原有最优识别结果和最优识别结果中的每个字对应的中间识别结果 ；词组库包括词组及其对应的中间识别结果 ；相似性匹配即计算词组库中的词组对应的中间识别结果与字混淆网络中的中间识别结果的相似程度，用相似概率值表示该相似程度，相似概率值大于零的词组为当前识别语句中可能再次出现的词 ；所述词组信息包括词组本身、词组的相似概率值以及词组在最优识别结果中的对应位置 ；语句中的重复出现词是指语句中的某个词在以前的语句中出现过，其在当前语句的再次出现叫做重复出现词，对其识别错误称为重复出现词识别错误 ；除了重复出现词识别错误外，还有首次出现的词被识别错误的情况，这种识别错误叫做非重复出现词识别错误。

[0007] 可选的，步骤 (1) 中所述进行相似性匹配包括 ：计算词组及对应中间识别结果与

当前字混淆网络的相似概率值；和保留相似概率值大于零的词组；其中，所述计算方式为：

[0008] $$p(W,CN_j)=\prod_{i=1}^{num(W)}SIM(S_i,S'_{i+j})$$

[0009] $$SIM(C,C')=\frac{1}{2}\left[\frac{1}{N_1}\sum_{i=0}^{N_1-1}\delta(c_i,C')+\frac{1}{N_2}\sum_{i=0}^{N_2-1}\delta(c'_i,C)\right]$$

[0010] num(W) 表示词组 W 中字的个数，$p(W,CN_j)$ 表示词组对应中间识别结果与当前字混淆网络中第 j 列开始的 num(W) 列中间识别结果的相似概率，$SIM(S_i,S'_{i+j})$ 表示词组中第 i 个字所对应的中间识别结果与当前字混淆网络中第 i+j 列中间识别结果的相似性；

[0011] C 和 C′ 分别表示一列中间识别结果集合，N_1 和 N_2 分别表示 C 和 C′ 中字的个数；c_i 表示 C 中的第 i 个字，c'_i 表示 C′ 中的第 i 个字；$\delta(c_i, C')$ 表示如果在 C′ 集合中存在某个字的读音与字 c_i 的读音相同，其值等于 1，否则等于 0。

[0012] 可选的，词组中的每个字对应当前字混淆网络中的一列识别结果；步骤 (2) 中所述重新计算相似概率值包括：

[0013] 重新计算查找到词组中每个字的相似概率值；

[0014] 其中，重新计算查找到词组中每个字的相似概率值方式为：

[0015] $$p_{c_i}=\begin{cases}\lambda p'_{c_i}+(1-\lambda)q'(c_{loc+i,k}) & c_{loc+i,k}=c_i\\ \lambda p'_{c_i} & else\end{cases}$$

[0016] $q'(c_{loc+i,k})$ 表示第 i 个字所对应的第 loc+i 列识别结果中第 k 个字的原有识别概率值，$c_{loc+i,k}=c_i$ 表示第 i 个字所对应的第 loc+i 列识别结果中存在一个与之相同的字。

[0017] 可选的，步骤 (2) 中所述重新计算字识别概率值包括：

[0018] 重新计算每个对应列中的字识别概率值；

[0019] 其中，重新计算每个对应列中的字识别概率值的方式为：

[0020] $$q(c_{loc+i,j})=(1-\lambda)q'(c_{loc+i,j})+\lambda(1-p'_{c_i})q'(c_{loc+i,j})$$

[0021] p'_{c_i} 表示词组 w 中第 i 个字的概率值，p'_{c_i} 等于 p_w，$q'(c_{loc+i,j})$ 表示第 i 个字所对应的第 loc+i 列识别结果中第 j 个字的原有识别概率值，相应的 $q(c_{loc+i,j})$ 表示重新计算后的识别概率值，λ 表示词组 w 为重复出现词的权重。

[0022] 可选的，所述的自动修正方法还包括：(5) 通过在混淆网络中选择正确的字、或者通过键盘输入、或者通过手写输入来修正非重复出现词识别错误，以得到不再包含任何识别错误的已修正识别结果。

[0023] 可选的，所述的自动修正方法还包括：

[0024] (6) 挖掘已修正识别结果中所有的词组；和

[0025] (7) 存储或更新得到的词组到词组库。

[0026] 可选的，步骤 (6) 中所述挖掘已修正识别结果中所有的词组包括：

[0027] 计算已修正识别结果中每个字与相邻若干字组成词的概率值；

[0028] 选择组合概率最大的词组作为挖掘到的词组；

[0029] 其中，所述计算的方式为：

[0030] $$p(c_1,c_2,\cdots,c_k)=\sum_{i_1=1}^{num(S_{c_1})}\sum_{i_2=1}^{num(S_{c_2})}\cdots\sum_{i_k=1}^{num(S_{c_k})}p(c_{1,i_1},c_{2,i_2},\cdots,c_{k,i_k})$$

[0031] $p(c_1, c_2, \cdots, c_k)$ 表示已修正识别结果中第 1 个字与第 2 个到第 k 个字组成词的概率值，$num(S_{c_k})$表示已修正识别结果中第 k 个字对应中间识别结果列中字的个数，c_{k,i_k}表示已修正识别结果中第 k 个字所对应中间识别结果列中的第 i_k 个字，$p(c_{1,i_1},c_{2,i_2},\cdots,c_{k,i_k})$表示字混淆网络中字$c_{1,i_1}$与$c_{2,i_2}$到$c_{k,i_k}$的组合概率值。

[0032] 可选的，所述词组库为词组文件或词组数据库。

[0033] 与现有技术相比，优点在于：利用之前已修正的识别结果中的经验知识，自动修正当前识别语句中重复出现词的识别错误，从而提高识别错误的修正效率，加快识别错误修正速度。

附图说明

[0034] 图 1 是本发明一个实施例中汉语语音识别中字混淆网络的示意图；

[0035] 图 2 是本发明一个实施例中挖掘词组的方法流程图；

[0036] 图 3 是本发明一个实施例中挖掘到的词组示意图；

[0037] 图 4 是本发明一个实施例中利用已修正识别结果自动修正当前识别语句中重复出现词识别错误的方法流程图；

[0038] 图 5 是图 4 中步骤 401 的流程图。

[0039] 图 6 是图 4 中步骤 402 的流程图。

[0040] 图 7 是本发明另一个实施例中利用已修正识别结果自动修正当前识别语句中重复出现词识别错误的方法流程图；

[0041] 图 8 是本发明又一个实施例中利用已修正识别结果自动修正当前识别语句中重复出现词识别错误的方法流程图。

具体实施方式

[0042] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图，根据实施例对本发明进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0043] 在汉语语音识别中，待识别的语音内容基本都是围绕着某个主题展开的，因此某些与主题相关的关键词会在前后的多句语句中出现。由于上下文及每次发音的差异性，同一关键词在不同语句中可能会被多次识别错误，即使在第一次出现时识别正确，在后续出现时也可能会被识别错误。如果对于每个重复出现的关键词，利用其第一次出现时的已修正识别结果，系统能够自动地修正其后续重复出现时的识别错误，则可以大大提高识别错误修正的效率，从而使语音识别应用能够真正被大多数用户所接受。

[0044] 下面首先介绍语音识别的基本过程、结果及修正。

[0045] 语音识别技术，也被称为自动语音识别 (Automatic Speech Recognition，ASR)，其目标是将人类的语音中的词汇内容转换为计算机可读的输入，例如按键、二进制编码或者字符序列。

[0046] 在语音识别过程中，字混淆网络是所有可能识别结果的集合。在字混淆网络中，每

个字都有一个识别概率值（即识别过程中生成该字的得分占识别总得分的比值）用以表示该字为识别结果的可能性，每列中的所有字之间具有竞争性且识别概率之和等于 1，此外，每个字还具有与相邻若干列中字组成词的组合概率值。在字混淆网络中，每一列中的字都按照识别概率值从大到小的顺序排列，字混淆网络中的第一行称为最优识别结果，最优识别结果中的每个字对应一列中间识别结果。

[0047] 如图 1 所示，为本发明一个实施例中字混淆网络的示意图。其中语音输入为 :gǔ、lǎo、dē、dōng、fáng。其可能的识别结果（即字混淆网络）100 包括最优识别结果 101 和中间识别结果 102。最优识别结果 101 为 :古、老、的、东、防。中间识别结果 102 包括“古”、“老”、“的”、“东”、“防”分别对应的中间识别结果，其中，“古”的中间识别结果为 :顾、孤、故，“老”的中间识别结果为 :乐、了，“的”的中间识别结果为 :得，“东”的中间识别结果为 :洞、冬，“防”的中间识别结果为 :房、放。

[0048] 可以看出，上述识别的结果最后一个词“dōng、fáng”的自动识别有误，而且“fáng”供选择的识别结果中没有“方”，所以需要通过手动输入，纠正识别的错误。经过上述识别错误修正后，识别结果就不再包含任何识别错误，称为已修正识别结果。

[0049] 在上述语音识别及修正过程中，正确的识别结果以及用户手动修改并输入的信息是可以重用的。为了自动修正后续识别语句中重复出现词的识别错误，需要将正确识别结果及用户修改识别结果的相关信息以某种形式进行保存。

[0050] 发明人经过分析发现，正确识别结果及用户修改识别结果的相关信息一般以词组的形式存在。这些词组在以后的识别语句中可能会再次出现，成为重复出现词。为修正以后识别语句中的这些重复出现词识别错误，需把这些词组挖掘出来并保存。为叙述方便，下述实施例中如果没有另外说明，已修正识别结果包括正确识别结果及用户修改识别结果。

[0051] 发明人经过分析还发现，针对不同的用户，同一词组所对应的可能的识别结果是不同的，即中间识别结果不同。以上述“dōng、fáng”为例，该用户所讲的“东、方”对应的中间识别结果分别为“洞、冬”和“房、放”，这也是该用户与其他用户的区别和特点 ;对于该用户之后的语音识别过程，相同的语音或语音序列一旦出现，很可能将以相似的中间识别结果表现出来。

[0052] 所以，为了保存用户修改识别结果的相关信息，既要存储所述已修正识别结果中的词语（即词组），还要存储这些词语对应的中间识别结果。根据本发明一个实施例，已修正识别结果中的词组挖掘是通过计算已修正识别结果中相邻字之间组成词的概率值来实现的，并选择概率值最大的组合作为词组，保存在词组库中。根据本发明另一个实施例，已修正识别结果中的词组挖掘还可以通过现有技术中的汉语分词实现，并保存在词组库中。

[0053] 图 2 是本发明一个实施例中提供的挖掘已修正识别结果中词组的流程图。

[0054] 步骤 201 :计算已修正识别结果中每个字与相邻若干字组成词的概率值。在已修正识别结果中，每个字对应一列中间识别结果。每个字与相邻若干字组成词的概率值等于对应中间识别结果中字之间组合概率之和，计算公式为 :

[0055] $$p(c_1,c_2,\cdots,c_k)=\sum_{i_1=1}^{num(S_{c_1})}\sum_{i_2=1}^{num(S_{c_2})}\cdots\sum_{i_k=1}^{num(S_{c_k})}p(c_{1,i_1},c_{2,i_2},\cdots,c_{k,i_k})$$

[0056] 其中，$p(c_1,c_2,\cdots,c_k)$ 表示已修正识别结果中第 1 个字与第 2 个到第 k 个字组成词的概率值，$num(S_{c_k})$表示已修正识别结果中第 k 个字对应中间识别结果列中字的个数，c_{k,i_k}

表示已修正识别结果中第 k 个字所对应中间识别结果列中的第 i_k 个字，$p(c_{1,i_1},c_{2,i_2},\cdots,c_{k,i_k})$ 表示字混淆网络中字c_{1,i_1}与c_{2,i_2}到c_{k,i_k}的组合概率值（即识别过程中这些字作为一个整体的识别得分占整个识别得分的比值）。

[0057] 步骤 202：选择组合概率最大的词组作为挖掘到的词组。在已修正识别结果中，每个字可以和后续相邻的一个，两个，或多个字组成词。因此，选择概率值最大组合作为词组，即选取最优结果。

[0058] 步骤 203：存储或更新挖掘到的词组。若挖掘到词组已在词组库中存在，则将词组对应的中间识别结果更新到对应词组库中的中间识别结果（即将对应词组库中间识别结果中没有的字添加到对应词组库中间识别结果中），否则将词组及对应的中间识别结果存储到词组库中。

[0059] 步骤 204：跳到已挖掘词组最后一个字的下一个位置，判断是否已超出已修正识别结果的范围，若是则结束，否则跳到步骤 201 进行下一个词组挖掘。

[0060] 在本发明一个实施例中，上述挖掘结果，即词组库的内容如图 3 所示。本领域技术人员可以理解，图 3 所示词组库既可以通过文件方式实现，也可以通过数据库的方式实现。通过该信息进行的语音识别和对重复出现词识别错误的自动修正过程将通过下面的实施例详细描述。

[0061] 图 4 是本发明一个实施例中提供的汉语语音识别中重复出现词识别错误的自动修正方法的流程图。对于每一句语音经语音识别引擎识别后都会生成一个字混淆网络，本方法就是从字混淆网络开始的，具体步骤如下：

[0062] 步骤 401：重复出现词查找。对每句话经识别后得到的字混淆网络与词组库中的词组（即词）及中间识别结果进行相似性匹配。相似性匹配即计算词组库中的词组对应的中间识别结果与字混淆网络中的中间识别结果的相似程度，用相似概率值表示该相似程度。相似概率值大于零的词组为当前识别语句中可能再次出现的词，保留该词组、相似概率值及其在最优识别结果中的对应位置。本实施例中，对第一句话来说，识别结果中的每个词都是第一次出现且此时的词组库为空，所以重复出现词查找结果为空。

[0063] 步骤 402：重复出现词识别错误修正。根据重复出现词查找得到的所有词组信息，包括词组本身、词组匹配的相似概率值、以及词组对应最优识别结果中的位置，重新计算当前字混淆网络中识别字的概率值。根据新的概率值以及词组相似概率值，对字混淆网络和词组按照概率值大小排序，通过词组替换原有最优识别结果中识别错误来修正重复出现词识别错误。

[0064] 具体的，步骤 401 如图 5 所示，查找当前识别语句中重复出现词的过程包括：

[0065] 步骤 4011：计算词组及对应中间识别结果与当前字混淆网络的相似概率值。词组的表示如图 3 所示，每个词组都具有与之对应的中间识别结果，词组中的每个字对应一行中间识别结果。用 $S_i = \{c_1, c_2, c_3, \cdots, c_j\}$ 表示词组 W 中第 i 个字所对应的一行中间识别结果，其中 c_j 表示中间识别结果中的第 j 个字；用 $S'_i = \{c'_1, c'_2, c'_3, \cdots, c'_k\}$ 表示当前字混淆网络 CN 中最优识别结果的第 i 个字所对应的一列中间识别结果，同样 c'_k 表示中间识别结果中的第 k 个字。词组及对应中间识别结果与当前字混淆网络相似概率值计算公式为：

[0066] $$p(W,CN_j)=\prod_{i=1}^{num(W)}SIM(S_i,S'_{i+j})$$

[0067] 其中，num(W) 表示词组 W 中字的个数，p(W，CN_j) 表示词组对应中间识别结果与当前字混淆网络中第 j 列开始的 num(W) 列中间识别结果的相似概率，SIM(S_i，S'_{i+j}) 表示词组中第 i 个字所对应的中间识别结果与当前字混淆网络中第 i+j 列中间识别结果的相似性。

[0068] $$SIM(C,C')=\frac{1}{2}\left[\frac{1}{N_1}\sum_{i=0}^{N_1-1}\delta(c_i,C')+\frac{1}{N_2}\sum_{i=0}^{N_2-1}\delta(c'_i,C)\right]$$

[0069] 其中，C 和 C′ 分别表示一列中间识别结果集合，N_1 和 N_2 分别表示 C 和 C′ 中字的个数。c_i 表示 C 中的第 i 个字，c'_i 表示 C′ 中的第 i 个字。δ(c_i，C′) 表示如果在 C′ 集合中存在某个字的读音与字 c_i 的读音相同，其值等于 1，否则等于 0。

[0070] 步骤 4012：保留相似概率值大于零的词组，若 p(W，CN_j) 大于 0 表示词组 W 可能在当前语句中出现，即当前识别语句中存在重复出现词。重复出现词出现的位置为最优识别结果中的第 j 个字，p(W，CN_j) 值越大表示 W 出现的可能性越大。因此对于 p(W，CN_j) 大于零的词组，保留词组 W、相似概率值 p(W，CN_j) 及出现位置 j。

[0071] 步骤 4013：判断当前词组是否为词组库中的最后一个词组，若是则结束重复出现词查找，否则回到步骤 4011 进行下一个词组的相似性匹配。

[0072] 具体的，在步骤 402 中，对当前字混淆网络与词组库进行相似性匹配得到的词组并非一定是当前识别语句的重复出现词，因为该词组可能只是与当前识别语句中的某个词具有发音相似性，而非真正的重复出现词。因此，在进行重复出现词识别错误修正时不能简单地用查找到的词组替换对应位置处的最优识别结果。本实施例中，根据重复出现词查找得到的所有词组信息，包括词组本身、词组匹配的相似概率值、以及词组对应最优识别结果中的位置，重新计算当前字混淆网络中字的识别概率值，根据新的概率值以及词组相似概率值来修正重复出现词识别错误。

[0073] 对重复出现词查找得到的每个词组用一个三元组表示 WI = {w，p_w，loc}，w 表示词组本身，p_w 表示词组 w 与当前字混淆网络的相似概率，loc 表示词组对应当前混淆网络中的开始位置，用 num(w) 表示词组 w 中字的个数，词组 w 与当前字混淆网络中从第 loc 列开始的 num(w) 列识别结果相对应，词组中的每个字对应当前字混淆网络中的一列识别结果，为了使查找到的词组的相似概率值与字混淆网络中对应中间识别结果字识别概率值具有可比性，且满足归一化的特点，重新计算相似概率值和字识别概率值。重新计算每个对应列中的字识别概率值的公式为：

[0074] $$q(c_{loc+i,j})=(1-\lambda)q'(c_{loc+i,j})+\lambda(1-p'_{c_i})q'(c_{loc+i,j})$$

[0075] 其中，p'_{c_i}表示词组 w 中第 i 个字的概率值，p'_{c_i}等于 p_w，q′ ($c_{loc+i,j}$) 表示第 i 个字所对应的第 loc+i 列识别结果中第 j 个字的原有识别概率值，相应的 q($c_{loc+i,j}$) 表示重新计算后的识别概率值，λ 表示词组 w 为重复出现词的权重。

[0076] 重新计算查找到词组中每个字的相似概率值公式为：

[0077] $$p_{c_i}=\begin{cases}\lambda p'_{c_i}+(1-\lambda)q'(c_{loc+i,k}) & c_{loc+i,k}=c_i\\ \lambda p'_{c_i} & else\end{cases}$$

[0078] 其中，p'_{c_i}、λ 同上，q′ $(c_{loc+i,k})$ 表示第 i 个字所对应的第 loc+i 列识别结果中第 k 个字的原有识别概率值，$c_{loc+i,k} = c_i$ 表示第 i 个字所对应的第 loc+i 列识别结果中存在一个与之相同的字。

[0079] 在完成概率值重新计算的基础上，将词组中的字以及该字对应当前字混淆网络列中的所有字一起按照概率值的大小从大到小排序。通过重新排序替换对应位置处的最优识别结果，从而修正对应的重复出现词识别错误。

[0080] 即如图 6 所示，步骤 402 进一步包括：

[0081] 步骤 4021，根据重复出现词查找得到的词组信息，重新计算相似概率值和字识别概率值；

[0082] 步骤 4022，根据新的概率值以及词组相似概率值，对字混淆网络和词组按照概率值大小排序；

[0083] 步骤 4023，使用排序结果替换字混淆网络的最优识别结果以及中间识别结果，从而修正对应的重复出现词识别错误。

[0084] 进一步的，在完成对最优识别结果中的重复出现词识别错误修正后，最优识别结果中可能还存在非重复出现词识别错误。由于最优识别结果中的每个字都对应一列中间识别结果，且中间识别结果与最优识别结果具有竞争性和发音相似性，因此对于某些非重复出现词识别错误可通过在对应中间识别结果中选择正确的字来修正。此外，还可以通过标识非重复出现词识别错误，然后用键盘输入或手写输入的方式来修正。

[0085] 图 7 是本发明一个实施例中提供的汉语语音识别中重复出现词识别错误的自动修正方法的流程图，所述方法包括：

[0086] 步骤 501：重复出现词查找；

[0087] 步骤 502：重复出现词识别错误修正；和

[0088] 步骤 503：非重复出现词识别错误修正。

[0089] 与上述实施例相比，其区别在于还包括步骤 503：非重复出现词识别错误修正。语句中的重复出现词是指语句中的某个词在以前的语句中出现过，其在当前语句的再次出现叫做重复出现词。在当前识别语句中，除了重复出现词识别错误外，还有首次出现的词被识别错误的情况，这种识别错误叫做非重复出现词识别错误。对于非重复出现词识别错误，通过在混淆网络中选择正确的字来修正识别错误，或者通过键盘输入，手写输入的方法来修正识别错误。经过非重复出现词识别错误修正后，识别结果就不再包含任何识别错误，成为已修正识别结果。

[0090] 进一步的，词组库中的词组可以动态生成、更新，而不需要事先准备好包含重复出现词组的词组库。

[0091] 图 8 是本发明一个实施例中提供的汉语语音识别中重复出现词识别错误的自动修正方法的流程图，所述方法包括：

[0092] 步骤 601：重复出现词查找；

[0093] 步骤 602：重复出现词识别错误修正；

[0094] 步骤 603：非重复出现词识别错误修正；

[0095] 步骤 604：词组挖掘；和

[0096] 步骤 605：存储或更新词组到词组库；

[0097] 与上述实施例相比，其区别在于还包括步骤 604 和 605。

[0098] 其中，步骤 604 ：词组挖掘。挖掘已修正识别结果中所有的词组，这些词组在以后的识别语句中可能会再次出现。已修正识别结果中的词组挖掘是通过计算已修正识别结果中相邻字之间组成词的概率值来实现的，对于每个字都选择概率值最大的组合作为词组。词组挖掘的具体步骤为上述步骤 201 ～ 204。

[0099] 步骤 605 ：存储或更新词组到词组库。将当前已修正识别结果中挖掘到的所有词组存储到词组库中，当词组在词组库中已存在时，则只需要更新词组所对应的中间识别结果，当词组库中不存在该词组时，则将词组以及其对应字混淆网络中的中间识别结果存储到词组库中。

[0100] 应用上述步骤 601 ～ 605，假设两句先后发出的语音对应的文本内容分别为 ：“修整遮盖胶带和色条”和“用胶带遮盖下围板”。在对第一句语音识别后得到的字混淆网络为 ：

[0101] 修 正 这 个 小 百 和 词 条

[0102] 就 诊 者 的 教 派 货 色 票

[0103] 纠 准 着 该 较 大 科 学 跳

[0104] 珍 胶 白

[0105] 带

[0106] 其中，第一句语音识别的最优识别结果为“修正这个小百和词条”，其中每个字都对应一列中间识别结果。由于第一句话中的每个词都是第一次出现，且此时词组库为空，因此对第一句语句字混淆网络的重复出现词查找为空。直接跳到非重复出现词识别错误修正，对最优识别结果中的第二个字“正”、第三个字“这”、第四个字“个”的识别错误通过键盘或手写输入“整”、“遮”、“盖”来修正，对最优识别结果中的第五个字“小”、第六个字“百”、第八个字“词”的识别错误通过在其对应的中间识别结果中选择“胶”、“带”、“色”来修正。在完成非重复出现词识别错误修正后，此时的最优识别结果“修整遮盖胶带和色条”为已修正识别结果。对已修正识别结果中的词组进行挖掘并存储，挖掘的结果如表 1 所示。

[0107] 表 1

[0108]

修整	就 纠
	诊 准 珍
遮盖	者 着
	的 该
胶带	教 较 胶
	派 大 白 带
色条	色 学
	票 跳

[0110] 在完成对第一句识别并修正后，对第二句话进行识别得到字混淆网络 ：

[0111] 有 小 的 这 个 小 礼 拜

[0112] 用 叫 在 着 的 下 对 白

[0113] 中 交 得 知 在 明 待

[0114] 教 派 者 该 李

[0115] 之

[0116] 其中，第二句话的最优识别结果为“有小的这个小礼拜”，每个字对应一列中间识别结果。计算词组库中词组与字混淆网络的相似概率值，词组“遮盖”与字混淆网络的相似概率值大于零，对应字混淆网络的第四列和第五列，词组“胶带”与字混淆网络的相似概率值大于零，对应字混淆网络的第二列和第三列。重新计算词组“遮盖”、“胶带”和它们所对应字混淆网络列中字的概率值，按照概率值的大小排序，排序后的结果为：

[0117] 有 胶 带 遮 盖 小 礼 拜

[0118] 用 小 的 这 个 下 对 白

[0119] 中 叫 在 着 的 明 待

[0120] 交 得 知 在 李

[0121] 教 派 者 该

[0122] 之

[0123] 通过重复出现词识别错误修正后，修正了最优识别结果中的第二个字“小”、第三个字“的”、第四个字“这”、第五个字“个”的识别错误。对于剩下的非重复出现词识别错误“有”、“小”、“礼”、“拜”通过从中间识别结果中选择候选或终端输入的方法修正，修正后的最优识别结果为“用胶带遮盖下围板”。对第二句已修正的识别结果挖掘词组，挖掘到的词组有“胶带”、“遮盖”和“围板”，其中“胶带”和“遮盖”已存在于词组库中，对于这两个词组只需更新对应的中间识别结果。存储和更新后的结果如表 2 所示。

[0124] 表 2

[0125]

修整	就 纠
	诊 准 珍
遮盖	者 着 这 知 之
	的 该 个 在
胶带	教 较 胶 叫 小 交
	派 大 白 带 的 在 得
色条	色 学
	票 跳
围板	对 明 李
	白 待

[0127] 应该注意到并理解，在不脱离后附的权利要求所要求的本发明的精神和范围的情况下，能够对上述详细描述的本发明做出各种修改和改进。因此，要求保护的技术方案的范围不受所给出的任何特定示范教导的限制。

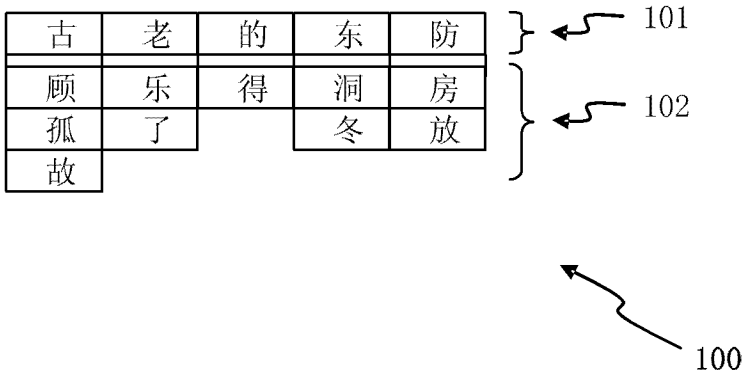

古
老
的
东
防
顾
乐
得
洞
房
孤
了
冬
放
故
101
102
100

图 1

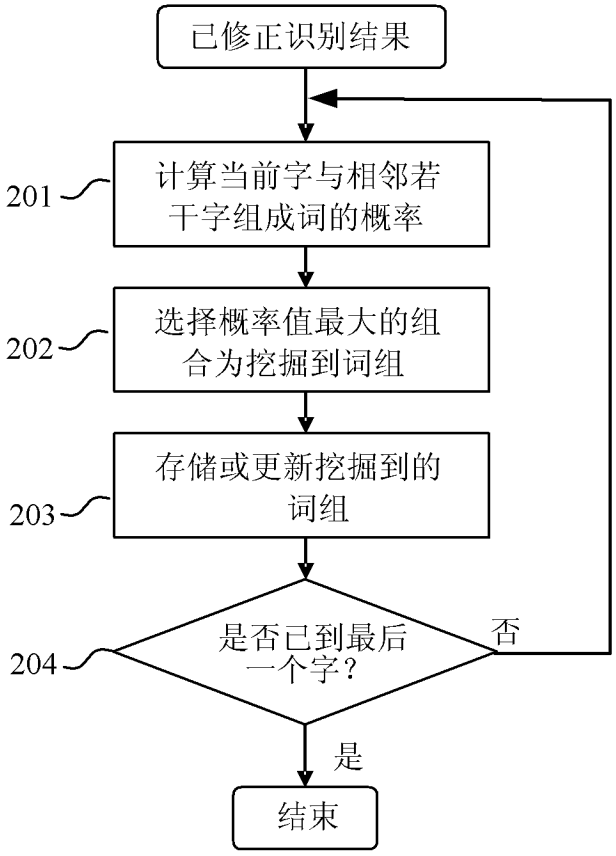

已修正识别结果
计算当前字与相邻若干字组成词的概率
201
选择概率值最大的组合为挖掘到词组
202
存储或更新挖掘到的词组
203
是否已到最后一个字？
204
否
是
结束

图 2

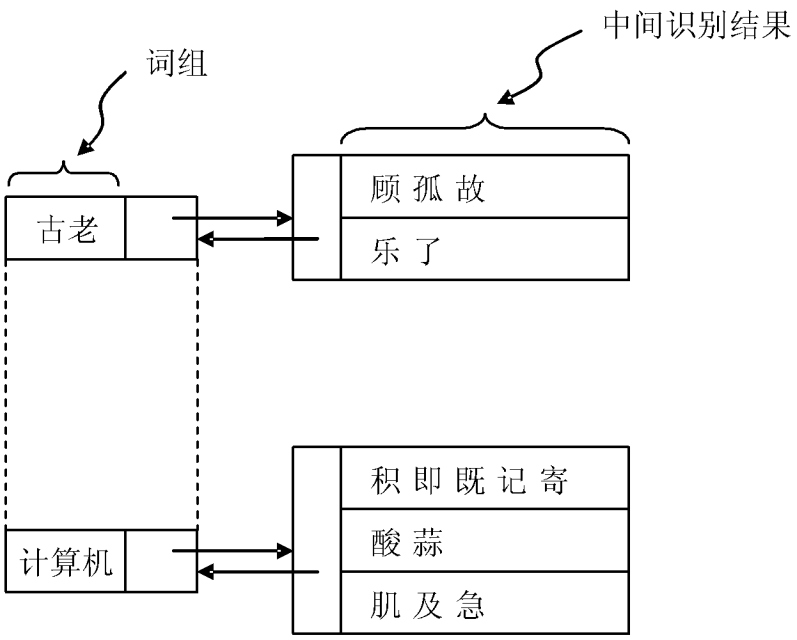

词组
中间识别结果
古老
顾 孤 故
乐 了
计算机
积 即 既 记 寄
酸 蒜
肌 及 急

图 3

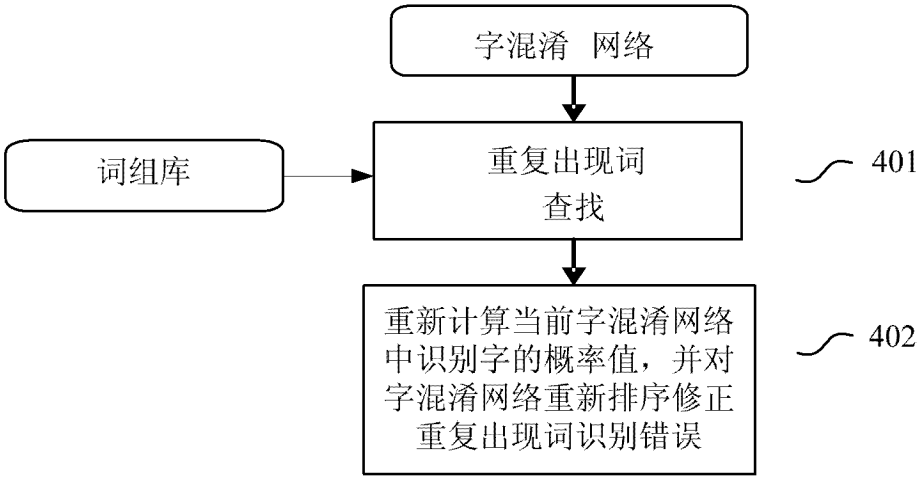

字混淆 网络
词组库
重复出现词
查找
401
重新计算当前字混淆网络
中识别字的概率值，并对
字混淆网络重新排序修正
重复出现词识别错误
402

图 4

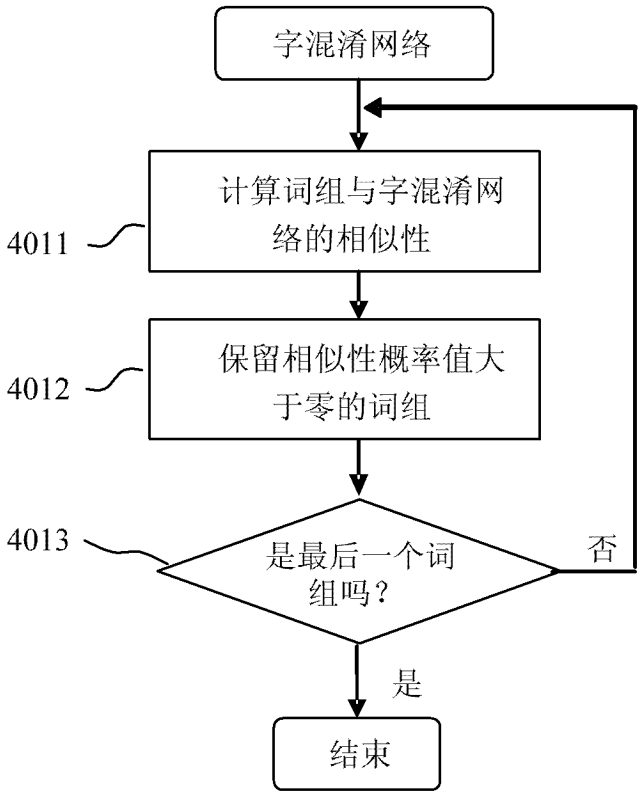

字混淆网络
计算词组与字混淆网络的相似性
4011
保留相似性概率值大于零的词组
4012
是最后一个词组吗？
4013
否
是
结束

图 5

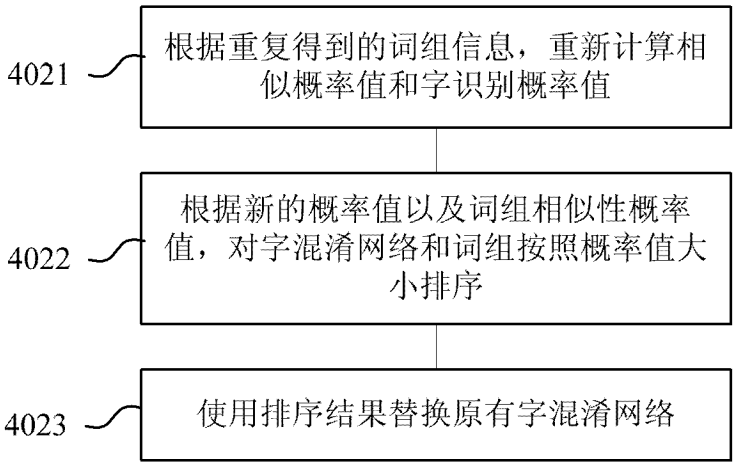

根据重复得到的词组信息，重新计算相似概率值和字识别概率值
4021
根据新的概率值以及词组相似性概率值，对字混淆网络和词组按照概率值大小排序
4022
使用排序结果替换原有字混淆网络
4023

图 6

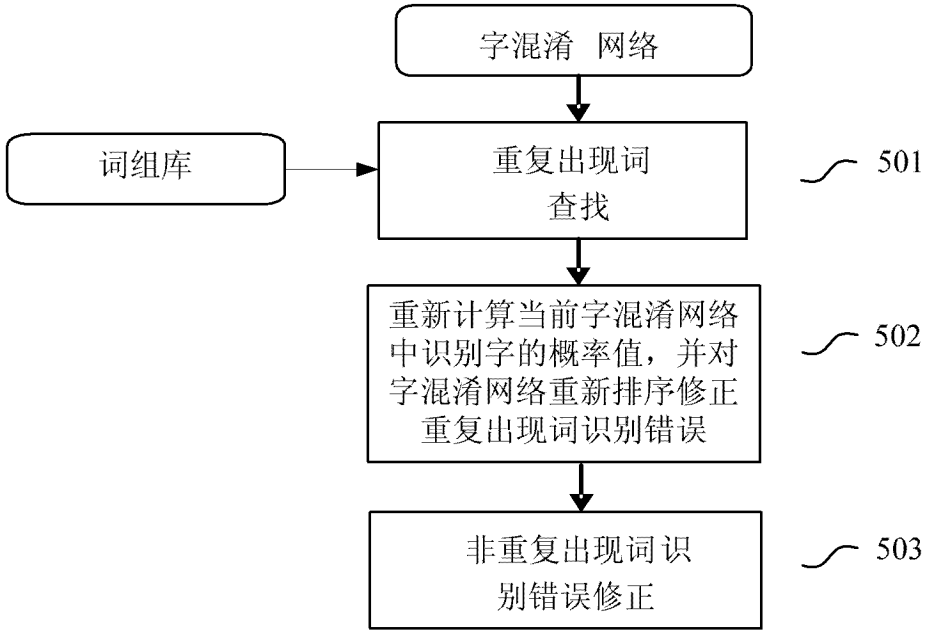

字混淆 网络
词组库
重复出现词
查找
501
重新计算当前字混淆网络
中识别字的概率值，并对
字混淆网络重新排序修正
重复出现词识别错误
502
非重复出现词识
别错误修正
503

图 7

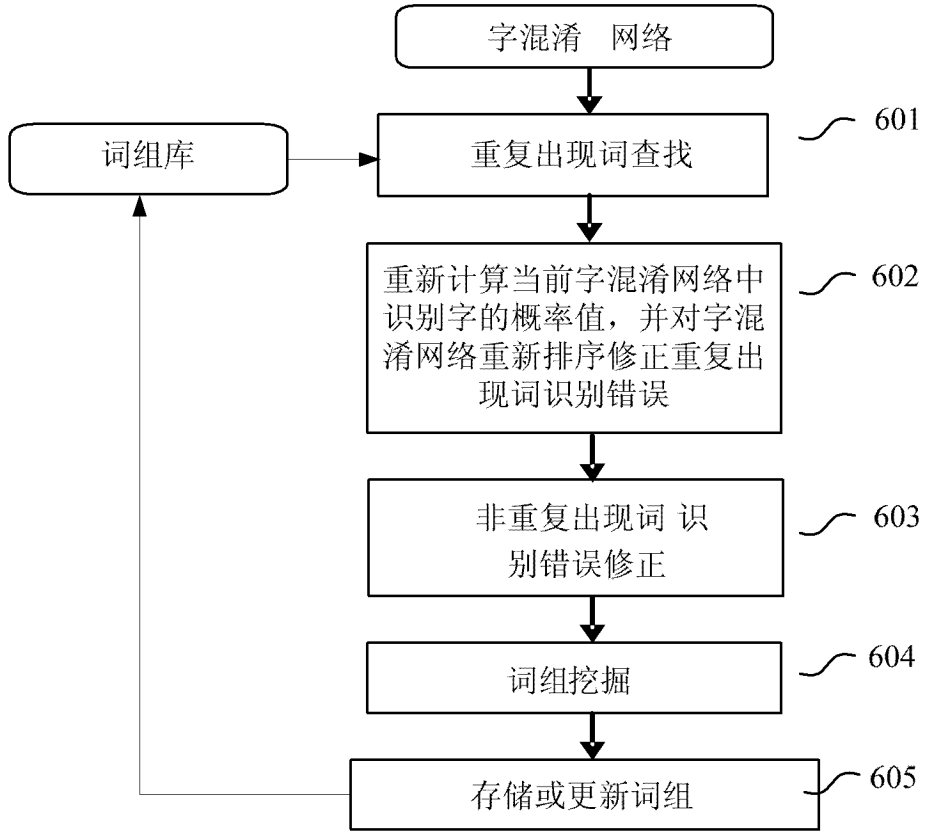

字混淆 网络
词组库
重复出现词查找
601
重新计算当前字混淆网络中
识别字的概率值，并对字混
淆网络重新排序修正重复出
现词识别错误
602
非重复出现词 识
别错误修正
603
词组挖掘
604
存储或更新词组
605

图 8