(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 1667700 B
(45)授权公告日 2010.10.06

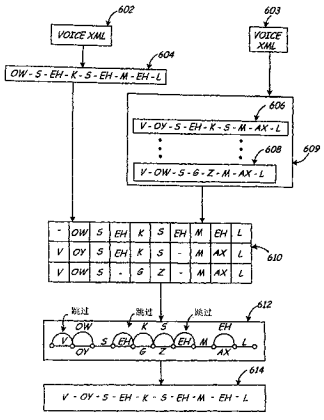
(21)申请号 200510054537.7
(22)申请日 2005.03.10
(30)优先权数据
10/796,921 2004.03.10 US
(73)专利权人 微软公司
地址 美国华盛顿州
(72)发明人 M·-Y·黄
(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 陈斌
(51)Int.Cl.
G10L 15/06(2006.01)
G10L 15/28(2006.01)
G10L 15/00(2006.01)
(56)对比文件
US 5031113 A,1991.07.09, 全文.
US 5857099 A, 全文.
审查员 隋欣
权利要求书 3 页 说明书 11 页 附图 7 页

(54)发明名称

把字的语音或声学描述、发音添加到语音识别词典的方法

(57)摘要

方法和计算机可读介质把字的文本和该字的用户发音转换成语音描述，以添加到语音识别词典中。开始，产生至少两个的多个可能语音描述。通过解码代表该字的用户发音的语音信号形成一个语音描述。从该字的文本产生至少一个其它语音描述。包括基于语言和基于文本的语音描述的多个可能序列，基于其对用户发音的对应物，在单个图表中进行对准和计分。然后选择最高分的语音描述作为语音识别词典中的条目。



CN 1667700 B

1. 一种用于把字的语音描述添加到语音识别词典中的方法，包括：

在对语言信号进行解码之前使用互信息来产生一类音节单元集合，以标识类音节单元的序列；

产生字的基于语言的语音描述，而无需引用所述字的文本；

基于所述字的文本产生所述字的基于文本的语音描述；

在逐音素基础上对准所述基于语言的语音描述和所述基于文本的语音描述，以形成单个图表；以及

从所述单个图表选择语音描述，其中所选择的语音描述既包括基于语言的语音描述的语音单元，也包括基于文本的语音描述的语音单元。

2. 如权利要求 1 所述的方法，其特征在于，还包括基于该字的用户发音产生所述基于语言的语音描述。

3. 如权利要求 2 所述的方法，其特征在于，还包括对代表所述字的用户发音的语言信号进行解码以产生所述字的基于语言的语音描述。

4. 如权利要求 3 所述的方法，其特征在于，对语言信号进行解码包括从所述语言信号识别类音节单元的序列。

5. 如权利要求 1 所述的方法，其特征在于，使用互信息来产生类音节单元包括：

计算训练字典中字分段单元对的互信息值；

基于所述互信息值来选择字分段单元对；以及

把所选择的字分段单元对合并成类音节单元。

6. 如权利要求 2 所述的方法，其特征在于，产生所述基于文本的语音描述包括使用字母 - 声音规则。

7. 如权利要求 1 所述的方法，其特征在于，从所述单个图表中选择语音描述包括：比较语言样本和所述单个图表中语音单元的声学模型。

8. 一种用于把字的语音发音添加到语音识别词典中的方法，包括：

接收字的文本，该字的语音发音要被添加到语音识别词典中；

接收由某人发音所述字而产生的语言信号的表示；

把所述字的文本转换成语音单元的至少一个基于文本的语音序列；

从所述语言信号的表示中产生语音单元的基于语言的语音序列；

把所述至少一个基于文本的语音序列的语音单元和所述基于语言的语音序列的语音单元置入搜索结构中，所述搜索结构允许所述基于文本的语音序列的语音单元和所述基于语言的语音序列的语音单元之间的转移；以及

从所述搜索结构中选择语音发音，其中所选择的语音发音既包括所述基于语言的语音序列的语音单元，也包括所述基于文本的语音序列的语音单元。

9. 如权利要求 8 所述的方法，其特征在于，把语音单元置入搜索结构中包括：对准所述基于语言的语音序列和所述至少一个基于文本的语音序列，以标识彼此作为选择对象的语音单元。

10. 如权利要求 9 所述的方法，其特征在于，对准所述基于语言的语音序列和所述至少一个基于文本的语音序列包括：计算这两种语音序列之间的最小距离。

11. 如权利要求 9 所述的方法，其特征在于，选择所述语音发音部分地基于语音单元的

声学模型和所述语言信号的表示之间的比较。

12. 如权利要求 8 所述的方法，其特征在于，产生语音单元的基于语言的语音序列包括：

产生语音单元的多个可能的语音序列；

使用至少一个模型以为每个可能的语音序列产生概率得分；以及

选择具有最高分的可能的语音序列作为语音单元的所述基于语言的语音序列。

13. 如权利要求 12 所述的方法，其特征在于，使用至少一个模型包括使用声学模型和语言模型。

14. 如权利要求 13 所述的方法，其特征在于，使用语言模型包括使用基于类音节单元的语言模型。

15. 如权利要求 12 所述的方法，其特征在于，选择语音发音包括基于至少一个模型对通过所述搜索结构的路径计分。

16. 如权利要求 15 所述的方法，其特征在于，所述至少一个模型包括声学模型。

17. 如权利要求 9 所述的方法，其特征在于，所述搜索结构包含在所述基于文本的语音序列和所述基于语言的语音序列中都发现的语音单元的单个路径。

18. 一种用于把字的声学描述添加到语音识别词典中的方法，其特征在于，所述方法包括：

基于所述字的文本产生基于文本的语音描述；

产生基于语言的语音描述，而无需引用所述字的文本；

在结构中对准所述基于文本的语音描述和所述基于语言的语音描述，所述结构包括表示语音单元的路径，至少一个来自所述基于文本的语音描述的语音单元的路径被连接到来自所述基于语言的语音描述的语音单元的路径；

通过所述结构选择路径序列；以及

基于所选定的路径序列产生所述字的声学描述，其中所述声学描述既包括所述基于语言的语音描述的语音单元，也包括所述基于文本的语音描述的语音单元。

19. 如权利要求 18 所述的方法，其特征在于，选择路径序列包括产生所述结构中路径的得分。

20. 如权利要求 19 所述的方法，其特征在于，产生路径的得分包括比较字的用户发音和所述结构中的语音单元的模型。

21. 如权利要求 19 所述的方法，其特征在于，还包括基于所述字的文本产生多个基于文本的语音描述。

22. 如权利要求 21 所述的方法，其特征在于，产生基于语言的语音描述包括对包括所述字的用户发音的语言信号进行解码。

23. 如权利要求 22 所述的方法，其特征在于，对语言信号进行解码包括使用类音节单元的语言模型。

24. 如权利要求 23 所述的方法，其特征在于，还包括通过以下步骤构建类音节单元的所述语言模型：

计算训练字典中类音节单元对的互信息值；

基于所述互信息值选择类音节单元对；以及

移除所选择的对，并用新的类音节单元来置换训练字典中的被移除的所选择的对。

25. 如权利要求 24 所述的方法，其特征在于，还包括：

重新计算所述训练字典中剩下的类音节单元对的互信息值；

基于经重新计算的互信息值选择新的类音节单元对；以及

移除所述新的类音节单元对，并用第二个新的类音节单元来置换训练字典中的所述新的类音节单元对。

26. 如权利要求 25 所述的方法，其特征在于，还包括使用所述训练字典以产生类音节单元的语言模型。

把字的语音或声学描述、发音添加到语音识别词典的方法

技术领域

[0001] 本发明涉及语音识别，尤其涉及通过组合基于语言和基于文本的语音描述来改进新字发音以产生发音。

背景技术

[0002] 本发明涉及语音识别，尤其涉及通过组合基于语言和基于文本的语音描述来改进新字发音以产生发音。

[0003] 在语音识别中，人类的语音被转换成文本。为了执行该转换，语音识别系统标识可产生语音符号的最可能的声学单元序列。为了减少必须执行的计算量，大多数系统限制对用感兴趣的语言代表文字的声学单元序列的搜索。

[0004] 声学单元序列和文字之间的映射存储在至少一个词典中（有时称为字典）。不管词典有多大，语音信号中的某些字将会在词典之外。语音识别系统不能识别这些词汇表外(OOV)的字，因为该系统不知道它们的存在。例如，有时在口述期间，用户会发现系统未识别一口述字。这会发生，因为系统对特定字定义有与用户发音不同的发音，即用户带有外国口音地发音该字。有时，字根本就不在词汇表中。相反，识别系统会被强迫去识别其它字来替换词汇表外字，从而导致了识别错误。

[0005] 在过去的语音识别系统中，通过提供字的拼写以及带有用户声音的字的声学样本或发音，用户可添加语音识别系统未识别的字。

[0006] 使用字母－声音规则，字的拼写被转换成一组语音描述。将输入字存储为上下文无关语法(CFG)的仅有条目。通过把语音样本应用到语音描述中的音素(phone)的声学模型中，可对它计分。每个语音描述的总分包括语言模型得分。在CFG中，语言模型概率等于1比CFG中每个节点上的分支数。然而，由于输入字是CFG中仅有的条目，仅有从起始节点起的一个分支(CFG中仅有的另一节点是结束节点)。结果，出于字母－声音规则的任何语音描述总是具有语言模型概率为1。

[0007] 在单独的解码路径中，通过标识类音节单元序列将语音样本转换为语音描述，而该类音节单元序列基于类音节单元中音素的声学模型和类音节单元n-字母语言模型来提供最佳组合的声学和语言模型总分。

[0008] 然后比较通过字母－声音CFG标识的语音序列总分以及通过类音节单元n-字母解码标识的类音节单元的最可能序列总分。具有最高总分的语音序列被选为字的语音序列。

[0009] 因而，在现有技术系统中，以两个单独并行路径来执行字母－声音解码和类音节单元解码。这因众多原因而不太理想。

[0010] 首先，由于两个路径未使用共同的语言模型，两个路径之间的总分不能总进行具有意义的比较。特别地，由于CFG的语言模型总是提供概率1，字母－声音语音描述的总分通常将比类音节单元描述更高，这取决于n-字母语言模型的概率通常会远远小于1。（类音节单元的语言模型概率是10^{-4}的数量级。）

[0011] 正因为这个，即使当声学样本更匹配于来自类音节单元路径的语音描述时，现有技术系统仍然倾向于来自字母－声音规则的语音序列。

[0012] 第二种精确性问题在产生诸如“voicexml”的组合字发音时发生。重要的是要注意在现有技术系统中 CFG 路径和 n- 字母音节路径彼此无关。因而，类似“voicexml”的组合字可导致发音错误，因为选定的发音必须是 CFG 发音或者是 n- 字母音节发音。然而，与 CFG 引擎一起使用的字母－声音 (LTS) 规则趋于在类似“voice”的相对可预测的字上执行得较好，而在像“xml”的正确发音几乎与其如何拼写无关的不可预测字上执行得较差。

[0013] 相反，n- 字母音节模型通常在产生类似“xml”的字的发音时就执行得相当好，因为它试图捕捉声学样本中的任何声音或字节序列而与拼写无关。然而对类似“voice”的可预测字它执行得就不如 CFG 引擎了。

[0014] 基于这些原因，如果用两个单独路径估算来自两个解码系统的语音描述，从由例如带有首字母缩拼词的可预测字，诸如“voicexml”，组合的组合字可导致发音错误。

[0015] 用于改进诸如“voicexml”的组合字发音的语音识别系统将具有重要的功用。

发明内容

[0016] 方法和计算机可读介质把字的文本和该字的用户发音转换成语音描述，以添加到语音识别词典中。开始，产生至少两个的多个可能语音描述。通过解码代表字的用户发音的语音信号形成一个语音描述。从该字的文本产生至少一个其它语音描述。对准包括基于语言和基于文本的语音描述的多个可能序列以产生发音图表。然后通过再次使用用户的发音语音，再次对发音图表计分。然后选择最高分的语音描述作为语音识别词典中的条目。

[0017] 本发明的一方面是使用类音节单元 (SLU) 来把语音发音解码成语音描述。类音节单元通常比单个音素大但比字小。本发明提供用于使用不需要语言专用语言学规则的基于互信息的数据驱动方法来定义这些类音节单元的装置。可在语音解码过程中构建和使用基于这些类音节单元的语言模型。

[0018] 本发明的另一方面使用户能输入与对应于拼写的典型发音很不相同的字的可听发音。例如，当输入英文字的文本时，能可听地对外文字发音。在本发明的该方面下，可从词典检索添加到词典中的新字语音描述，并将其转换成包括例如英文字的外文翻译的可听信号。

[0019] 本发明提供了一种用于把字的语音描述添加到语音识别词典中的方法，包括 ：产生字的基于语言的语音描述，而无需引用所述字的文本 ；基于所述字的文本产生所述字的基于文本的语音描述 ；在逐音素基础上对准所述基于语言的语音描述和所述基于文本的语音描述，以形成单个图表 ；以及从所述单个图表选择语音描述，其中所选择的语音描述既包括基于语言的语音描述的语音单元，也包括基于文本的语音描述的语音单元。

[0020] 如上所述的方法，还包括基于该字的用户发音产生所述基于语言的语音描述。

[0021] 如上所述的方法，还包括对代表所述字的用户发音的语言信号进行解码以产生所述字的基于语言的语音描述。

[0022] 如上所述的方法，对语言信号进行解码包括从所述语言信号识别类音节单元的序列。

[0023] 如上所述的方法，还包括在对语言信号进行解码之前使用互信息来产生一类音节

单元集合，以标识类音节单元的序列。

[0024] 如上所述的方法，使用互信息来产生类音节单元包括：计算训练字典中字分段单元对的互信息值；基于所述互信息值来选择字分段单元对；以及把所选择的字分段单元对合并成类音节单元。

[0025] 如上所述的方法，产生所述基于文本的语音描述包括使用字母－声音规则。

[0026] 如上所述的方法，从所述单个图表中选择语音描述包括：比较语言样本和所述单个图表中语音单元的声学模型。

[0027] 本发明还提供了一种用于把字的语音发音添加到语音识别词典中的方法，包括：接收字的文本，该字的语音发音要被添加到语音识别词典中；接收由某人发音所述字而产生的语言信号的表示；把所述字的文本转换成语音单元的至少一个基于文本的语音序列；从所述语言信号的表示中产生语音单元的基于语言的语音序列；把所述至少一个基于文本的语音序列的语音单元和所述基于语言的语音序列的语音单元置入搜索结构中，所述搜索结构允许所述基于文本的语音序列的语音单元和所述基于语言的语音序列的语音单元之间的转移；以及从所述搜索结构中选择语音发音，其中所选择的语音发音既包括所述基于语言的语音序列的语音单元，也包括所述基于文本的语音序列的语音单元。

[0028] 如上所述的方法，把语音单元置入搜索结构中包括：对准所述基于语言的语音序列和所述至少一个基于文本的语音序列，以标识彼此作为选择对象的语音单元。

[0029] 如上所述的方法，对准所述基于语言的语音序列和所述至少一个基于文本的语音序列包括：计算这两种语音序列之间的最小距离。

[0030] 如上所述的方法，选择所述语音发音部分地基于语音单元的声学模型和所述语言信号的表示之间的比较。

[0031] 如上所述的方法，产生语音单元的基于语言的语音序列包括：产生语音单元的多个可能的语音序列；使用至少一个模型以为每个可能的语音序列产生概率得分；以及选择具有最高分的可能的语音序列作为语音单元的所述基于语言的语音序列。

[0032] 如上所述的方法，使用至少一个模型包括使用声学模型和语言模型。

[0033] 如上所述的方法，使用语言模型包括使用基于类音节单元的语言模型。

[0034] 如上所述的方法，选择语音发音包括基于至少一个模型对通过所述搜索结构的路径计分。

[0035] 如上所述的方法，所述至少一个模型包括声学模型。

[0036] 如上所述的方法，所述搜索结构包含在所述基于文本的语音序列和所述基于语言的语音序列中都发现的语音单元的单个路径。

[0037] 本发明还提供了一种用于把字的声学描述添加到语音识别词典中的方法，其特征在于，所述方法包括：基于所述字的文本产生基于文本的语音描述；产生基于语言的语音描述，而无需引用所述字的文本；在结构中对准所述基于文本的语音描述和所述基于语言的语音描述，所述结构包括表示语音单元的路径，至少一个来自所述基于文本的语音描述的语音单元的路径被连接到来自所述基于语言的语音描述的语音单元的路径；通过所述结构选择路径序列；以及基于所选定的路径序列产生所述字的声学描述，其中所述声学描述既包括所述基于语言的语音描述的语音单元，也包括所述基于文本的语音描述的语音单元。

[0038] 如上所述的方法，选择路径序列包括产生所述结构中路径的得分。

[0039] 如上所述的方法，产生路径的得分包括比较字的用户发音和所述结构中的语音单元的模型。

[0040] 如上所述的方法，还包括基于所述字的文本产生多个基于文本的语音描述。

[0041] 如上所述的方法，产生基于语言的语音描述包括对包括所述字的用户发音的语言信号进行解码。

[0042] 如上所述的方法，对语言信号进行解码包括使用类音节单元的语言模型。

[0043] 如上所述的方法，还包括通过以下步骤构建类音节单元的所述语言模型：计算训练字典中类音节单元对的互信息值；基于所述互信息值选择类音节单元对；以及移除所选择的对，并用新的类音节单元来置换训练字典中的被移除的所选择的对。

[0044] 如上所述的方法，还包括：重新计算所述训练字典中剩下的类音节单元对的互信息值；基于经重新计算的互信息值选择新的类音节单元对；以及移除所述新的类音节单元对，并用第二个新的类音节单元来置换训练字典中的所述新的类音节单元对。

[0045] 如上所述的方法，还包括使用所述训练字典以产生类音节单元的语言模型。

附图说明

[0046] 图 1 是本发明可在其中实现的一般计算环境框图。

[0047] 图 2 是本发明可在其中实现的一般移动计算环境框图。

[0048] 图 3 是本发明中语音识别系统的框图。

[0049] 图 4 是本发明一实施例的词典更新组件的框图。

[0050] 图 5 是本发明中把字添加到语音识别词典的方法的流程图。

[0051] 图 6 是示出本发明对特定字的实现的流程图。

[0052] 图 7 是构建一类音节单元集的流程图。

具体实施方式

[0053] 图 1 示出了本发明可在其上实现的适当计算系统环境 100 的示例。该计算系统环境 100 仅是适当计算环境的一个示例，并非旨在提出本发明使用或功能性范围的任何限制。计算环境 100 也不应被解释为对示例性操作环境 100 中所示的任一组件或其组合有任何依赖性或任何需求。

[0054] 本发明也可在很多其它通用或专用计算系统环境或配置中使用。适于本发明使用的众所周知的计算系统、环境、和 / 或配置的示例包括，但不限于，个人计算机、服务器计算机、手持式或膝上型装置、多处理器系统、基于微处理器的系统、机顶盒、可编程电器消费品、网络 PC、迷你计算机、大型机、电话系统、包括任一种以上系统或装置的分布式计算环境等等。

[0055] 本发明可以在计算机可执行指令的一般上下文中进行说明，诸如由计算机执行的程序模块。一般而言，程序模块包括执行具体任务或实现具体抽象数据类型的例程、程序、对象、组件、数据结构等等。本发明还可在任务由经通信网络链接的远程处理装置执行的分布式计算环境中实践。在分布式计算环境中，程序模块可置于包括存储器存储装置的本地和远程计算机存储介质。

[0056] 参照图1，实现本发明的示例性系统包括计算机110形式的通用计算装置。计算机110的组件可包括，但不限于，处理单元120、系统存储器130以及把包括系统存储器在内的各种系统组件耦合到处理单元120的系统总线121。系统总线121可能是若干总线结构类型中的任何一种，包括存储器总线或存储器控制器、外围总线、以及使用多种总线结构的任一种的本地总线。作为示例，而非限制，这些结构包括工业标准结构(ISA)总线、微信道结构(MSA)总线、扩展ISA(EISA)总线、视频电子标准协会(VESA)局部总线和也称为Mezzanine总线的外围部件互连(PCI)总线。

[0057] 计算机110通常包括各种计算机可读介质。计算机可读介质可以是能被计算机110访问的任何可用介质，并包括易失性和非易失性介质、可移动和不可移动介质。作为示例，而非限制，计算机可读介质可包括计算机存储介质和通信介质。计算机存储介质包括以任何方法或技术实现、用于存储诸如计算机可读指令、数据结构、程序模块或其它数据等信息的易失性和非易失性介质、可移动和不可移动介质。计算机存储介质包括但不限于RAM、ROM、EEPROM、闪存或其它存储器技术、CD-ROM、数字化视频光盘(DVD)或其它光学存储技术、磁卡、磁带、磁盘存储或其它磁性存储装置、或任何其它可用于存储所需信息并可由计算机110访问的介质。通信介质通常包括诸如载波或其它传输机制的已调制数据信号中的计算机可读指令、数据结构、程序模块、或其它数据，且包括任何信息输送介质。术语“调制数据信号”意指以信息中的信息编码方式设置或改变其一个或多个特征的信号。作为示例，而非限制，通信介质包括诸如有线网络或直线连接的有线介质，和诸如声学、射频、红外线和其它无线介质的无线介质。以上任何介质的组合也应包括在计算机可读介质的范围中。

[0058] 系统存储器130包括诸如只读存储器(ROM)131和随机存取存储器(RAM)132的易失性和/或非易失性存储器形式的计算机可读介质。包含有助于计算机110如起动时在元件间传送信息的基本例程的基本输入/输出系统(BIOS)133通常存储在ROM 131中。RAM 132通常包含可被处理单元120立即访问和/或现时操作的数据和/或程序模块。作为示例，而非限制，图1示出了操作系统134、应用程序135、其它程序模块136、和程序数据137。

[0059] 计算机110还可包括其它可移动/不可移动、易失性/非易失性计算机存储介质。作为示例，图1图示了读取和写入不可移动、非易失性磁性介质的硬盘驱动器141，读取和写入可移动、非易失性磁盘152的磁盘驱动器151，读取和写入可移动、非易失性光盘156，诸如CD-ROM或其它光学介质的光盘驱动器155。其它也用在示例性计算环境中的可移动/不可移动、易失性/非易失性计算机存储介质包括，但不限于，如磁带、闪存卡、数字化视频光盘、数字化录像带、固态RAM、固态ROM等等。硬盘驱动器141通常通过诸如接口140的不可移动存储器接口与系统总线121连接，而磁盘驱动器151和光盘驱动器155通常通过诸如接口150的可移动存储器接口与系统总线121连接。

[0060] 如上所述并如图1所示的盘驱动器及其相关联的计算机存储介质为计算机110提供计算机可读指令、数据结构、程序模块、和其它数据的存储。在图1中，例如，硬盘驱动器141被示为存储操作系统144、应用程序145、其它程序模块146、和程序数据147。注意这些组件可以与操作系统134、应用程序135、其它程序模块136、和程序数据137相同或不同。在此给予操作系统144、应用程序145、其它程序模块146、和程序数据147的数字不同至少说明他们是不同的复制件。

[0061] 用户可通过输入装置如键盘162、话筒163和诸如鼠标、跟踪球或触摸板等定位装

置 161 向计算机 110 输入命令和信息。其它输入装置（未示出）可包括话筒、游戏杆、游戏垫、卫星接收器、扫描仪、无线电接收器、或电视或广播视频接收器等等。这些和其它输入装置常常通过与系统总线耦合的用户输入接口 160 与处理单元 120 相连，但也可通过诸如并行端口、游戏端口或通用串行总线 (USB) 的其它接口连接。监视器 191 或其它类型的显示装置也可通过诸如视频接口 190 的接口与系统总线 121 相连。除了监视器，计算机还可包括诸如扬声器 197 和打印机 196 的其它输出装置，它们通过输出外围接口 195 相连。

[0062] 计算机 110 可以在使用与一台或多台远程计算机，诸如远程计算机 180 的逻辑连接的网络化环境中运行。远程计算机 180 可以是个人计算机、服务器、路由器、网络 PC、对等装置或其它普通网络节点，而且通常包括上述与个人计算机 110 相关的许多或全部组件，尽管在图 1 中仅图示了存储器存储装置 181。图 1 中所描绘的逻辑连接包括局域网 (LAN)171 和广域网 (WAN)173，但也可包括其它网络。这样的网络化环境在办公室、企业范围计算机网络、企业内部互联网和因特网上是常见的。

[0063] 当用于 LAN 网络化环境中时，计算机 110 通过网络接口或适配器 170 与局域网 171 连接。当用于 WAN 网络化环境中时，计算机 110 通常包括调制解调器 172 或其它用于在广域网 173，诸如因特网中建立通信的装置。可以是内置式或外置式的调制解调器 172 与系统总线 121 通过用户输入接口 160 或其它适当机制连接。在网络化环境中，与计算机 110 相关的程序模块或其一部分可存储在远程存储器存储装置中。作为示例，而非限制，图 1 示出了驻留于远程计算机 180 中的远程应用程序 185。应当理解，所示网络连接是示例性的，且其它用于在计算机间建立通信链路的技术也可以使用。

[0064] 图 2 是可选示例性计算环境的移动装置 200 的框图。移动装置 200 包括微处理器 202、存储器 204、输入 / 输出 (I/O) 组件 206、以及用于与远程计算机或其它移动装置进行通信的通信接口 208。在一实施例中，前述组件经适当总线 210 耦合用于彼此通信。

[0065] 存储器 204 被实现为诸如随机存取存储器 (RAM) 带有电池备份模块（未示出）的非易失性电子存储器，从而当移动装置 200 的总电源关闭时存储在存储器 204 中的信息不会丢失。存储器 204 的一部分更适于被分配为用于程序执行的可寻址存储器，而存储器 204 的另一部分更适于用来存储，诸如模拟盘驱动器上的存储。

[0066] 存储器 204 包括操作系统 212、应用程序 214 以及典型存储器 216。在操作期间，操作系统 212 更适于由处理器 202 从存储器 204 上执行。在一优选实施例中，操作系统 212 是可从微软公司购买的WINDOWS® CE 品牌的操作系统。操作系统 212 更适于为移动装置设计，并实现可由应用程序 214 通过一组外露应用程序编程接口和方法利用的数据库特征。由应用程序 214 和操作系统 212 至少部分地响应于对外露应用程序编程接口和方法的调用，来维护对象存储器 216 中的对象。

[0067] 通信接口 208 代表使移动装置 200 能够发送和接收信息的多种装置和技术。这些装置包括有线和无线调制解调器、卫星接收器和广播调谐器（仅列举若干）。移动装置 200 还可直接与计算机连接以交换数据。这样，通信接口 208 可以是都能够传输流信息的红外线收发器或串行或并行通信连接。

[0068] 输入 / 输出组件 206 包括各种输入设备，诸如触摸感应屏幕、按钮、滚轴和话筒以及各种输出设备，诸如音频发生器、振动设备、和显示器。以上所列设备作为示例，且无需都在移动装置 200 上出现。另外，其它输入 / 输出设备可附于移动装置 200 或与其一体，在本

发明范围之内。

[0069] 图 3 提供了与本发明特别相关的语音识别模块的更详细框图。在图 3 中，如果需要可由话筒 300 把输入语音符号转换为电子信号。然后通过模拟 - 数字或 A/D 转换器把电子信号转换成电子信号。在若干实施例中，A/D 转换器 302 在 16kHz 对模拟信号取 16 比特的样本，因此创建每秒 32KB 的语音数据。

[0070] 向框架建构单元 304 提供数字式数据，该单元把数字式值分组成值的帧。在一实施例中，每个帧 25 毫秒长，且在前一帧开始之后 10 毫秒开始。

[0071] 向从数字信号中提取特征的特征提取器 304 提供数字式数据的帧。特征提取模块的示例包括用于执行线性预测编码 (LPC)、LPC 导出对数倒频谱、感应式线性预测 (PLP)、听觉模型特征提取、以及 Mel 频率对数倒频谱系数 (MFCC) 特征提取。注意，本发明并不限于这些特征提取模块，且可在本发明的上下文中使用其它模块。

[0072] 特征提取器 306 能每帧产生一个多维特征向量。特征向量中的维数和数值量取决于使用的特征提取类型。例如，Mel 频率对数倒频谱系数向量通常具有 12 个系数，加上一个代表总共 13 维的幂数的系数。在一实施例中，通过取 Mel 频率系数的一阶导数和二阶导数加上相对于时间的幂，特征向量可从 Mel 系数计算。因而，对于这种特征向量，每帧都与形成特征向量的 39 个数值相关联。

[0073] 在语音识别期间，由特征提取器 306 产生的特征向量流提供给解码器 308，该解码器基于特征向量流、系统词典 310、应用程序词典 312（如果有）、用户词典 314、语言模型 316、以及声学模型 318 来标识词的最可能或最相似序列。

[0074] 在大多数实施例中，声学模型 318 是由一组隐藏状态组成的隐藏马尔可夫模型，其中每个输入信号帧一个状态。每个状态具有描述匹配特定状态的输入特征向量可能的相关联概率分布组。在某些实施例中，概率的混合（通常为 10 个高斯概率）关联于每个状态。隐藏马尔可夫模型还包括用于在两个相邻模型状态之间转换、以及用于特定语言学单元状态之间转换的概率。对于本发明的不同实施例，语言学单元的尺寸可不同。例如，语言学单元可以是 senones、音素、双音素、三音素、音节、甚至整个字。

[0075] 系统词典 310 由对特定语言有效的语言学单元（通常为字或音节）列表组成。解码器 308 使用系统词典 310 以把其对可能语言学单元的搜索限制在那些确实是语言一部分的那些单元。系统词典 310 还包含发音信息（即从每个语言学单元映射到由声学模型 318 使用的一个声学单元序列）。可任选的应用词典 312 与系统词典 310 相似，除了应用词典 312 包含由特定应用添加的语言学单元，而系统词典 310 包含由语音识别系统提供的语言学单元。用户词典 314 也与系统词典 310 相似，除了用户词典 314 包含已由用户添加的语言学单元。在本发明中，提供了用于把新语言学单元添加到特别是用户词典 314 的方法和装置。

[0076] 语言模型 316 提供特定的语言学单元序列将出现在特定语言中的一组似然性或概率。在许多实施例中，语言模型 316 基于诸如北美商业新闻 (NAB) 的文本数据库，诸如在题为“CSR-III Text Language Model”宾州州立大学 1994 的文章有更详细的描述。语言模型 316 可以是上下文无关语法、诸如三字母 (trigram) 的统计学 n- 字母模型、或两者的组合。在一实施例中，语言模型 316 是紧密三字铭模型，它基于字序列的三字分段的组合概率确定该序列的概率。

[0077] 基于声学模型 318、语言模型 316、以及词典 310、312、314，解码器 308 从所有可能的语言学单元序列中标识最可能的语言学单元序列。该语言学单元序列代表语音信号的誊本。

[0078] 该誊本被提供给输出模型 320，它处理与把该誊本发送给一个或多个应用程序相关联的开销成本。在一实施例中，输出模块 320 与存在于图 3 语音识别引擎和一个或多个应用程序之间的中间层（如果有）进行通信。

[0079] 在本发明中，可通过在用户界面 321 上输入字的文本来把新字添加到用户字典 314 中。由 A/D 转换器 302、帧建构器 304 以及特征提取器 306 把发音字转换成特征向量。在添加字的过程期间，这些特征向量提供给字典更新单元 322 而不是解码器 308。更新单元 322 还从用户界面 321 接收新字的文本。基于特征向量和新字的文本，字典更新单元 322 通过以下进一步描述的过程来更新用户字典 314 和语言模型 316。

[0080] 图 4 提供了用于更新用户词典 314 和语言模型 316 的词典更新单元 322 的组件的框图。图 5 提供了由图 4 组件实现的用于更新用户词典 314 的方法的流程图。

[0081] 在步骤 502，用户通过对着话筒念字输入新字以产生用户提供声学样本 401。用户提供声学样本 401 如上所述被转换成提供给词典更新单元 322 的特征向量 403。特别地，将特征向量 403 提供给类音节单元 (SLU) 引擎 405 以在图 5 的步骤 504 产生由特征向量 403 代表的类音节单元的最可能序列。SLU 引擎 405 包括或访问 SLU 字典 409 和声学模型 318 以通常基于最高概率得分产生 SLU 的最可能序列。然后 SLU 引擎 403 把类音节单元的最可能序列转换成提供给对准模块 414 的语音单元序列。SLU 字典 409 将对应于以下的图 7 进行更详细描述。

[0082] 重要的是要注意在某些情形中用户对新字的发音与典型发音极为不同。例如，说话者可能通过代之以英文字的外文翻译来发音该英文字。例如这种特征将使得语音识别词典以一种语言存储字的文本或拼写，而用不同于第一种语言的第二语言存储语音描述。

[0083] 在步骤 506，用户输入新字的文本以产生用户提供文本样本 402。注意步骤 506 可在步骤 502 之前、之后、或者与之同时执行。用户提供文本样本 402 被提供给语法模块 404，它在步骤 508 将该文本转换成可能的基于文本的语音序列列表。尤其是语法模块 404 为用户提供文本样本 402 构建诸如上下文无关语法的语法。语法模块 404 包括或访问词典 406 以及字母 - 声音 (LTS) 引擎 408。语法模块 404 首先搜索包括系统词典 310、可任选应用词典 312、以及用户词典 314 的词典 406 来为用户提供文本样本 402（如果有）以检索可能的语音描述、发音、或序列。

[0084] LTS 引擎 408 把用户提供的文本样本 402 转换成一个或多个可能语音序列，特别是当没有在词典 406 中发现该字时。通过利用适于感兴趣的特定语言的发音规则集合 410 来执行该转换。在大多数实施例中，语音序列由一系列音素构建而成。在其它实施例中，该语音序列是三音素序列。语法模块 404 因而从词典 406 和 LTS 引擎 408 产生了一个或多个可能的基于文本语音序列 412。

[0085] 再看图 4，向对准模块 414 提供来自 SLU 引擎 405 的最佳语音序列 407 以及来自语法模块 404 的可能语音序列列表 402。在步骤 510，对准模块 414 以与用于计算例如来自置换错误、删除错误、以及插入错误的语音识别误差率的众所周知的对准模块和 / 或方法相似的方式来对准语音序列 404 和 412。在某些实施例中，使用两个序列字符串之间的最小

距离（例如正确的基准和识别假设）可执行对准。对准模块 414 产生经对准语音序列的列表、图表或表格。

[0086] 在步骤 511，对准模块 414 把经对准语音序列置入单个图表。在该过程期间，相互对准的同一语音单元在单一路径上进行组合。相互对准的相异语音单元则置于图表上的并行的备选路径上。

[0087] 该单个图表被提供给重新计分模块 416。在步骤 512，再次使用特征向量 403 以对由遍布该单个图表的路径所代表的语音单元的可能组合重新计分。在一实施例中，使用通过沿路径比较由用户对字的发音产生的特征向量 403 和存储在声学模型 318 中的每个语音单元的模型参数而产生的声学模型得分，重新计分模块 416 执行维特比（Viterbi）搜索以标识该图表中的最佳路径。该计分过程与由解码器 308 在语音识别期间执行的计分过程相似。

[0088] 得分选择和更新模块 418 选择单个图表中最高得分语音序列或路径。选中序列被提供以在步骤 514 更新用户词典 314，及在步骤 516 更新语言模型 316。

[0089] 图 6 示出了本发明如何处理或学习字的发音的示例。框 602 示出字“voicexml”的用户发音，而框 603 代表“voicexml”的输入文本。字“voicexml”用于说明本发明在如上所述产生组合字的发音中的优点。字“voicexml”的第一部分或“voice”是诸如图 4 中 LTS 引擎 408 的 LTS 引擎通常能精确处理的相对可预测字或字分段。然而，该字的第二部分“xml”是 LTS 引擎会有精确处理问题的不可预测或非典型字或首字母缩拼词。然而，诸如 SLU 引擎 405 的典型 SLU 引擎通常可很好地处理诸如“xml”的字或字分段，因为 SLU 引擎取决于用户的声学发音。

[0090] 框 604 示出诸如由图 4 中 SLU 引擎 405 和图 5 中步骤 504 产生的最可能语音序列。因而，字“voicexml”的声学或口语版的最佳发音如下：

[0091] ow-s-eh-k-s-eh-m-eh-l。

[0092] 在此情形中，用户没有对语音单元“v”发音或者 SLU 模型未较好地预测语音单元“v”。结果，在语音序列的开始去掉了可预期的语音单元“v”。

[0093] 在框 609 字“voicexml”的拼写或文本版本的可能语音序列 606 和 608 的列表由 LTS 引擎 408 产生，包括语音单元的以下序列：

[0094] v-oy-s-eh-k-s-m-ax-l，

[0095] v-ow-s-g-z-m-ax-l。

[0096] 由对准模块 414 在框 610 所示的对准结构中组合来自框 604 和 609 的语音序列。通常，使用动态规划以及基于给定各种对准下语音序列之间差异的成本函数来执行该对准。在框 610，经对准语音单元重新在同一竖直列中。注意某些栏具有标识没有语音单元与之相关联的空路径的“-”，意思是该栏是可任选的或可跳过的。

[0097] 框 612 示出由包括可从经对准结构形成的可能语音序列的经对准结构 610 构建的单个图表。框 612 表示语音单元可置于节点间路径上的搜索结构。在该结构中，在从 SLU 引擎标识的语音单元、基于语言的语音单元、以及由 LTS 引擎标识的语音单元、基于文本的语音单元之间允许进行转移。框 612 还示出选中路径可包括“跳过”，其中该路径中特定栏未包括语音单元。

[0098] 如上所述，使用字的用户发音和声学模型来选择语音序列或路径。框 614 示出根

据本发明的选中语音序列或路径，且提供如下：

[0099] v-oy-s-eh-k-s-eh-m-eh-l。

[0100] 注意最后路径用由 LTS 引擎预测的语音序列开始，而用由 SLU 引擎预测的语音序列结束。在现有技术中，这是不可能的。因而，本发明从结合了来自基于语言 SLU 引擎和基于文本 LTS 引擎的可能语音序列的单个图表中选择语音序列，以产生字的更精确发音。

[0101] 类音节单元 (SLU) 集

[0102] 图 7 示出了构建可用于本发明某些实施例中的类音节单元 (SLU)409 的字典或集合的方法。通常，图 7 的方法是有利的，因为它不需要语言专用语言学规则。因而，图 7 所示的方法可用于任何语言且要实现它相对便宜，因为它不需要其它方法，特别是语言学基于规则方法所必需的熟练语言学家。

[0103] 图 7 的方法采用了互信息 (MI) 以构建 SLU 集，并使用了类似于用于不同环境的在题为“Modeling Out-of-vocabulary Words For Robust Speech Recognition”Issam Bazzi 2000 年的博士论文中所述算法的算法。在本发明中，给定一大语音字典，例如约有 5 万或更多字的带有语音描述的训练字典，预定或有限尺寸（例如 1 万单元）的类音节单元集合得以构建。

[0104] 在框 702，起始的 SLU 集 S_0 等于音素集 $P = \{p_1, p_2, \ldots p_n\}$，通常为在英语语音识别系统中发现的 40 个音素，从而 $S_0 = \{s_1, s_2, \ldots s_m\} = \{p_1, p_2, \ldots p_n\}$，其中 m 和 n 是分别 SLU 和音索的数量，且开始时 $m = n$。

[0105] 使 (u_1, u_2) 为当前迭代中的任一对 SLU。在框 704，在字典条目中发现的语言学单元对 (u_1, u_2) 的互信息由以下等式进行计算。

[0106] $$MI(u_1,u_2)=\Pr(u_1,u_2)\log\frac{\Pr(u_1,u_2)}{\Pr(u_1)\Pr(u_2)}$$ 等式 1

[0107] 其中 $MI(u_1, u_2)$ 是类音节单元对 (u_1, u_2) 的互信息，$\Pr(u_1, u_2)$ 是 (u_1, u_2) 的联合概率，且 $\Pr(u_1)$ 和 $\Pr(u_2)$ 分别是 u_1 和 u_2 的单字母概率。

[0108] 使用以下等式来计算单字母概率 $\Pr(u_1)$ 和 $\Pr(u_2)$：

[0109] $$\Pr(u_1)=\frac{Count(u_1)}{Count(*)}$$ 等式 2

[0110] $$\Pr(u_2)=\frac{Count(u_2)}{Count(*)}$$ 等式 3

[0111] 其中 $Count(u_1)$ 和 $Count(u_2)$ 分别是在训练字典中发现类音节单元 u_1 和 u_2 的次数，而 Count(*) 是训练字典中类音节单元实例的总数。(u_1, u_2) 的联合概率可由以下等式计算：

[0112] $$\Pr(u_1,u_2)=\Pr(u_2|u_1)\Pr(u_1)$$ 等式 4

[0113] $$=\frac{Count(u_1,u_2)}{Count(u_1,*)}\frac{Count(u_1)}{Count(*)}$$

[0114] $$=\frac{Count(u_1,u_2)}{Count(*)}$$

[0115] 其中 $Count(u_1, u_2)$ 是 (u_1, u_2) 对在训练字典中一起（即相邻）出现的次数。

[0116] 在框 706，选择或标识具有最大互信息的对 (u_1, u_2)。在框 708，带有最多互信息的对 (u_1, u_2) 被并入新的更长的类音节单元 u_3。新类音节单元 u_3 替换或置换训练字典的字中

的对 (u_1, u_2)。

[0117] 在框 710，判定是否要中止迭代。在某些实施例中，可使用控制 LSU 最大长度的参数。例如，最大类音节单元长度可设定为 4 个音素。如果达到了选中长度，则中止对选中对的合并，而改为检查具有最大互信息的下一对。如果不再有其它对或者如果 SLU 的数量达到所需数量、或者最多互信息降到某阈值之下，图 7 的方法进行到框 712，其中 SLU 集 S 被输出。否则，方法返回到框 704，其中产生新 u_3 之后重新计算类音节单元的互信息，并重新计算受影响单元的单字母和双字母计数。在一实施例中，在每次迭代中仅并入一对类音节单元。然而，在其它实施例中，如果速度是关注对象，诸如在 Bazzi 的论文中，可在每次迭代中合并选定的对数（例如 50 对）。

[0118] 当图 7 的算法结束，输入或训练字典被分成最终的 SLU 集。然后可从分段字典中训练类音节单元 n- 字母，并以本发明实现之。已发现该数据驱动方法能比基于规则的按音节发音方法获得略好的精确度。然而，更重要的是，可不作代码改变在任何语言中使用该方法，因为它不需要语言专用语言学规则。

[0119] 尽管已参照特定实施例描述了本发明，本领域技术人员将连接可在形式和细节上作改变，不背离本发明的精神和范围。

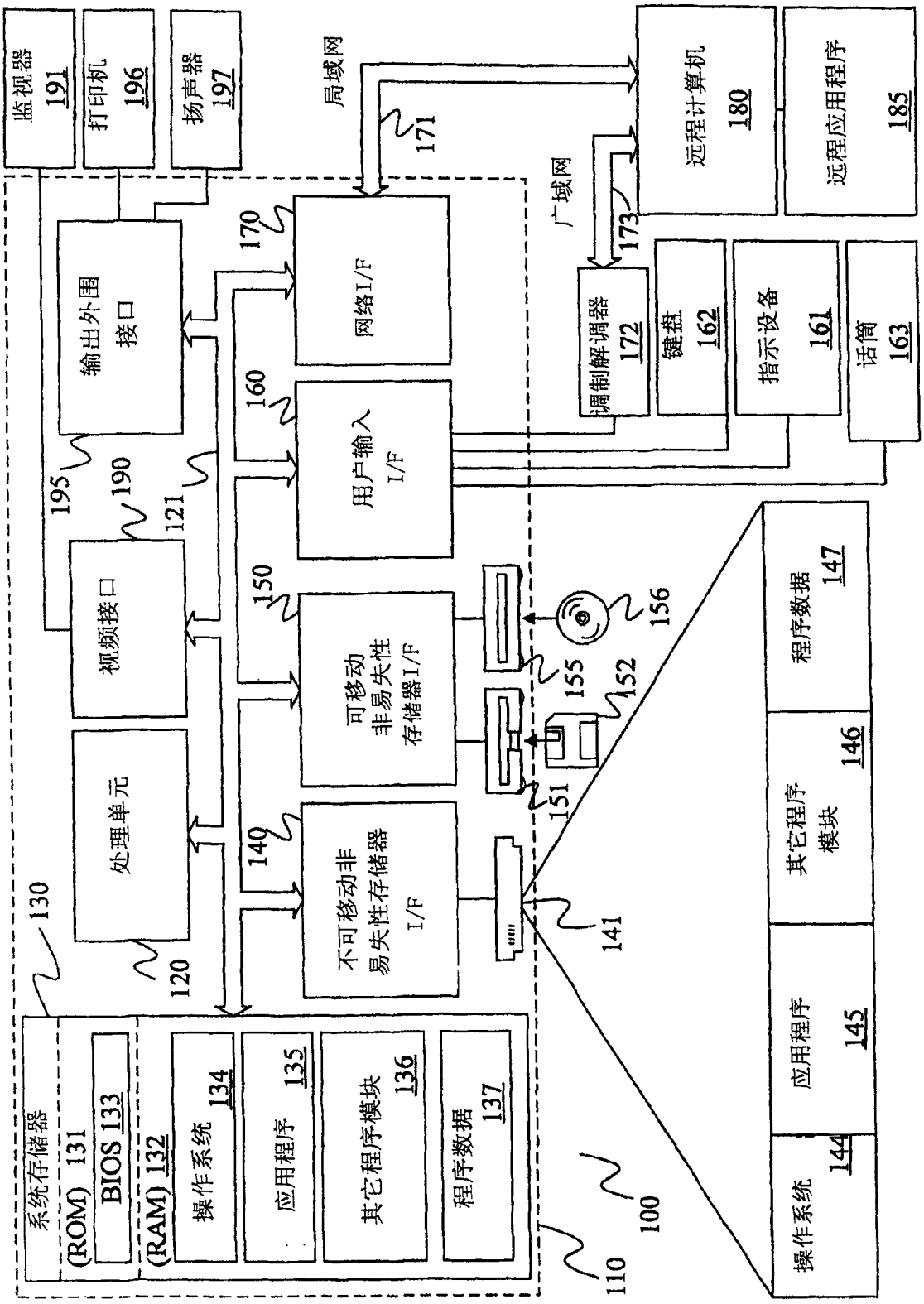

监视器 191
打印机 196
扬声器 197
局域网
171
远程计算机 180
远程应用程序 185
广域网
173
输出外围接口
170
网络I/F
调制解调器 172
键盘 162
指示设备 161
话筒 163
用户输入I/F
160
195
190
121
视频接口
150
可移动非易失性存储器I/F
156
155
152
151
处理单元
120
130
140
不可移动非易失性存储器I/F
141
程序数据 147
其它程序模块 146
应用程序 145
操作系统 144
系统存储器
(ROM) 131
BIOS 133
(RAM) 132
操作系统 134
应用程序 135
其它程序模块 136
程序数据 137
110
100

图 1

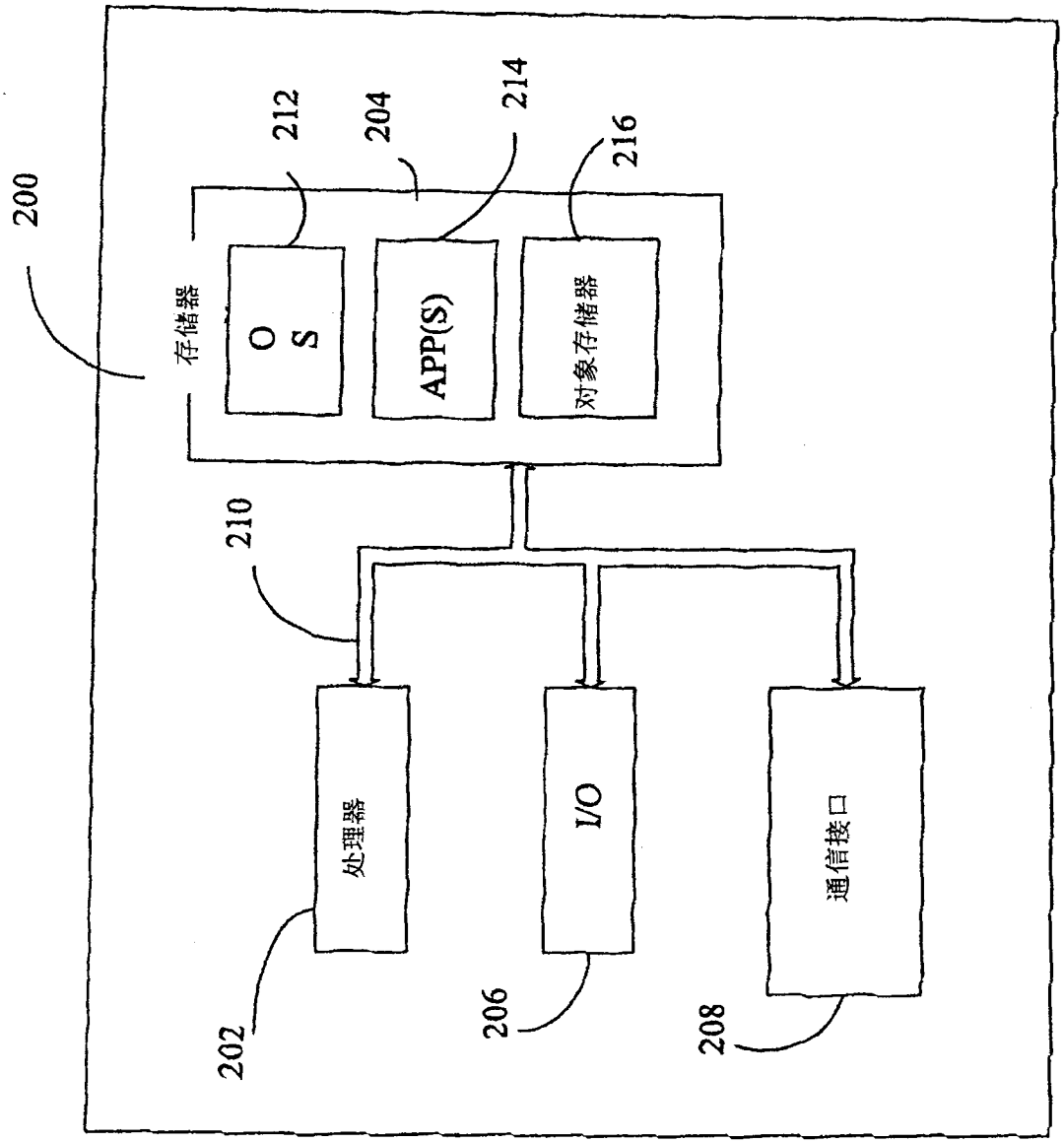

200
存储器
O
S
212
204
APP(S)
214
对象存储器
216
210
处理器
202
I/O
206
通信接口
208

图 2

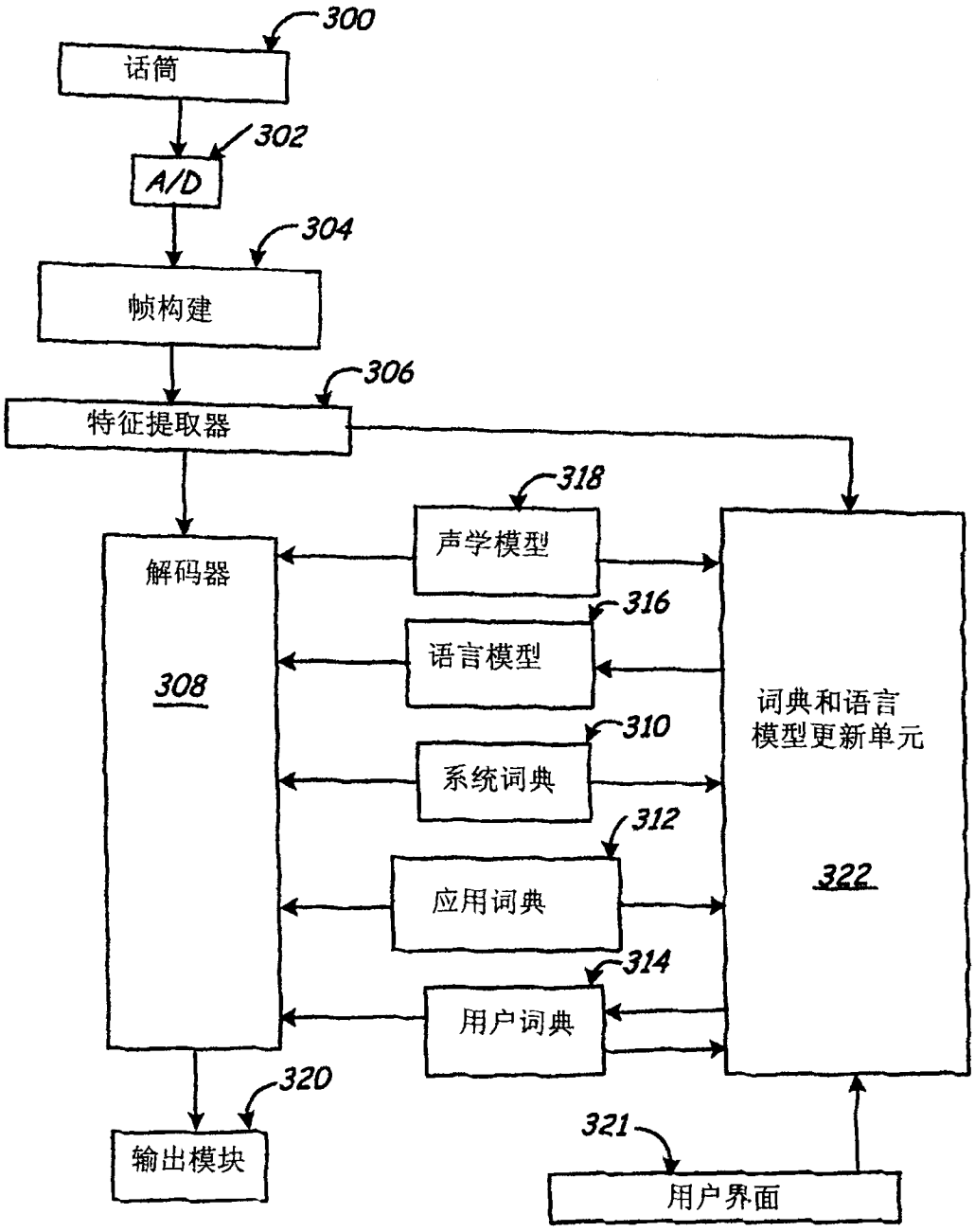

300
话筒
302
A/D
304
帧构建
306
特征提取器
318
声学模型
解码器
308
316
语言模型
310
系统词典
312
应用词典
314
用户词典
词典和语言
模型更新单元
322
320
输出模块
321
用户界面

图 3

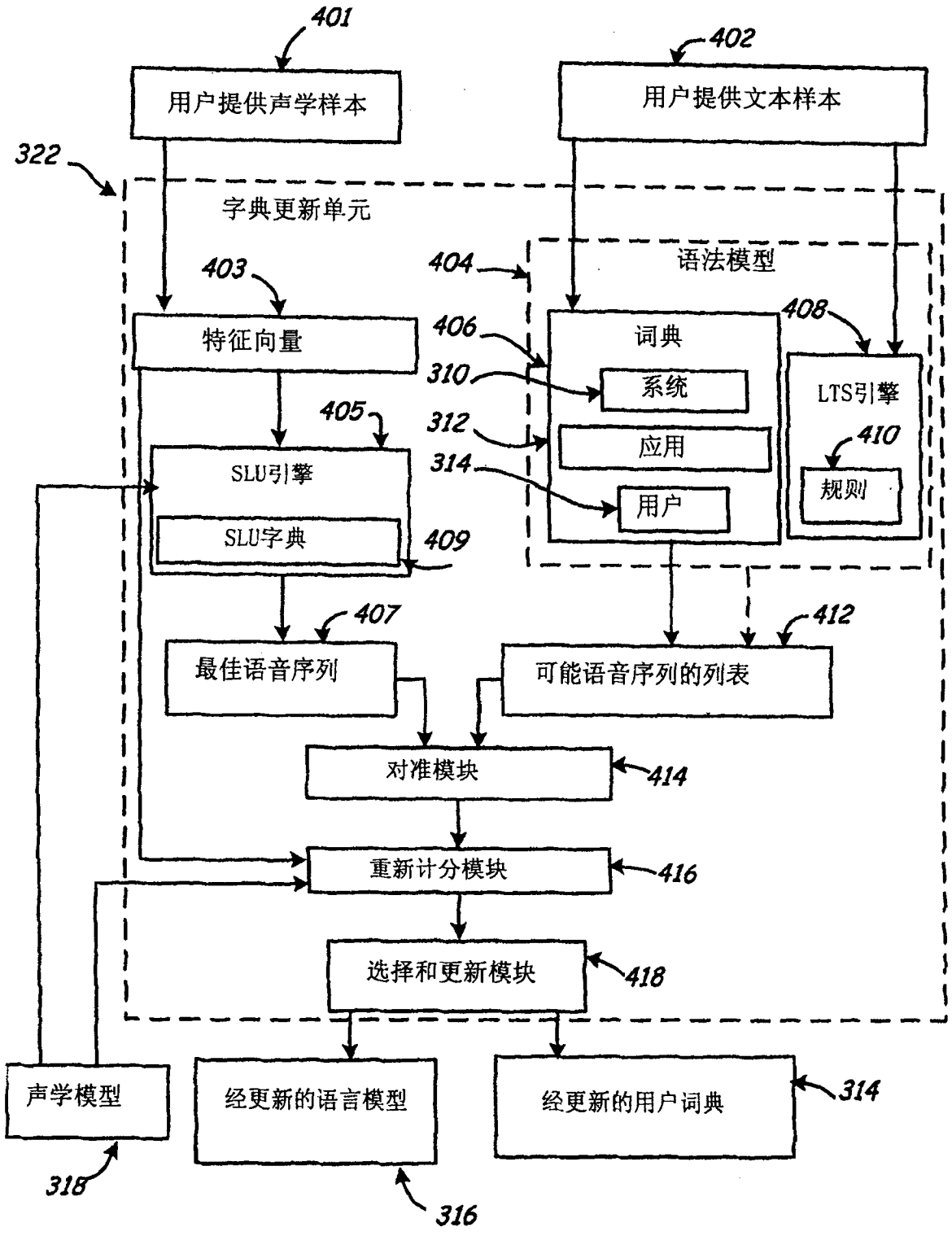

401
402
用户提供声学样本
用户提供文本样本
322
字典更新单元
403
404
语法模型
特征向量
406
词典
408
310
系统
LTS引擎
405
312
应用
410
SLU引擎
314
规则
用户
SLU字典
409
407
412
最佳语音序列
可能语音序列的列表
对准模块
414
重新计分模块
416
选择和更新模块
418
声学模型
经更新的语言模型
经更新的用户词典
314
318
316

图 4

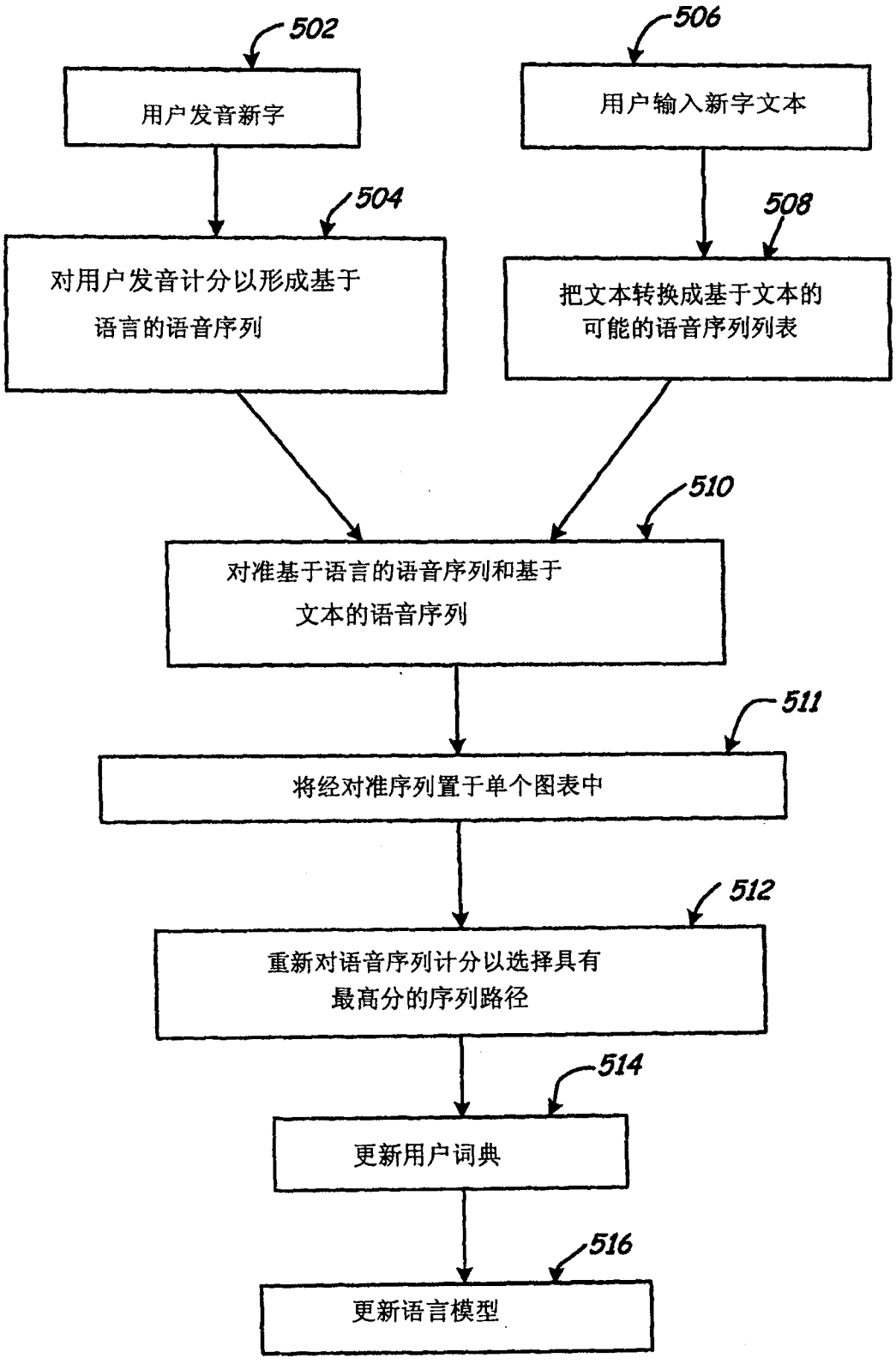

502
用户发音新字
506
用户输入新字文本
504
对用户发音计分以形成基于语言的语音序列
508
把文本转换成基于文本的可能的语音序列列表
510
对准基于语言的语音序列和基于文本的语音序列
511
将经对准序列置于单个图表中
512
重新对语音序列计分以选择具有最高分的序列路径
514
更新用户词典
516
更新语言模型

图 5

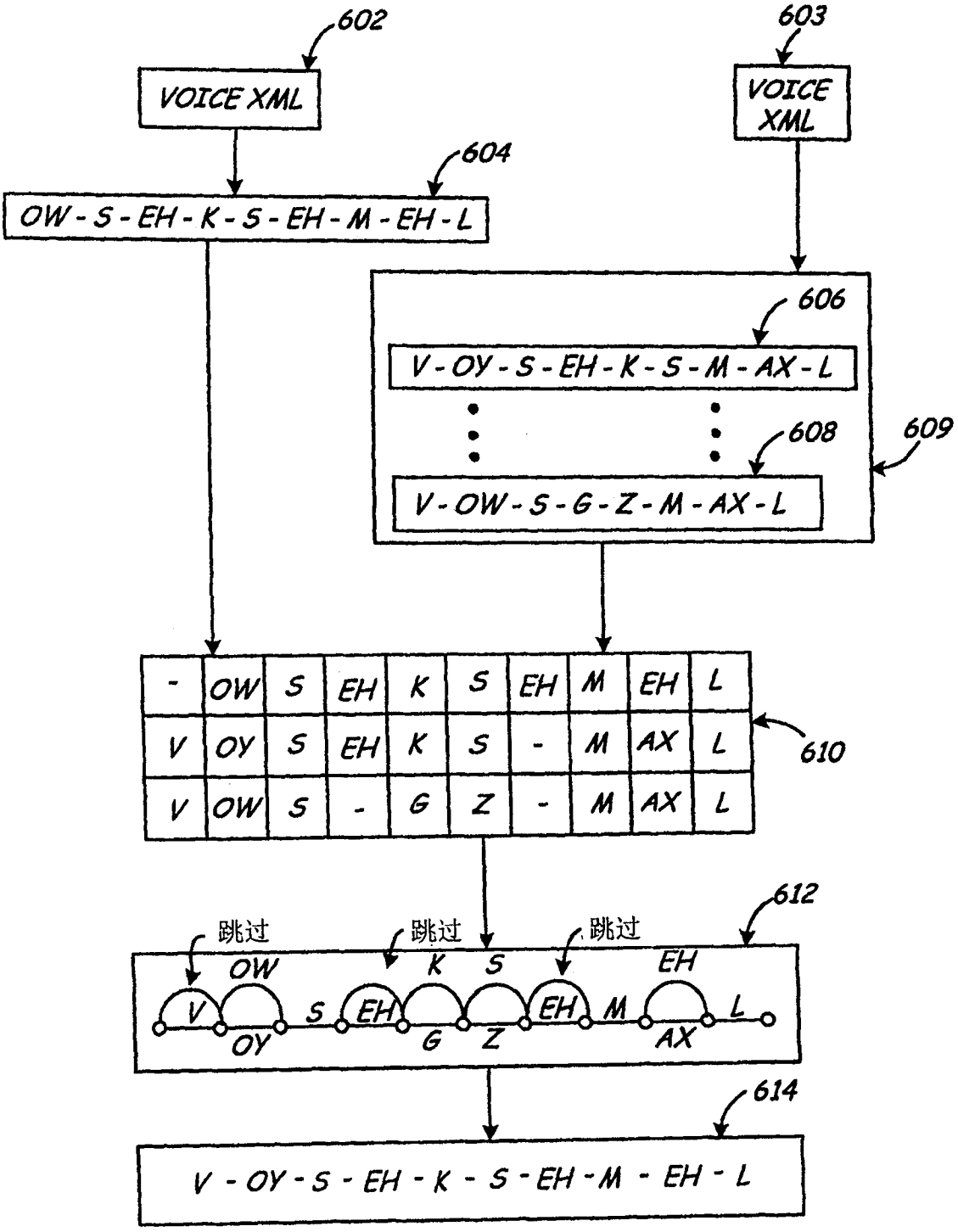

602
VOICE XML
603
VOICE XML
604
OW - S - EH - K - S - EH - M - EH - L
606
V - OY - S - EH - K - S - M - AX - L
608
609
V - OW - S - G - Z - M - AX - L
610
612
跳过
跳过
跳过
614
V - OY - S - EH - K - S - EH - M - EH - L

图 6

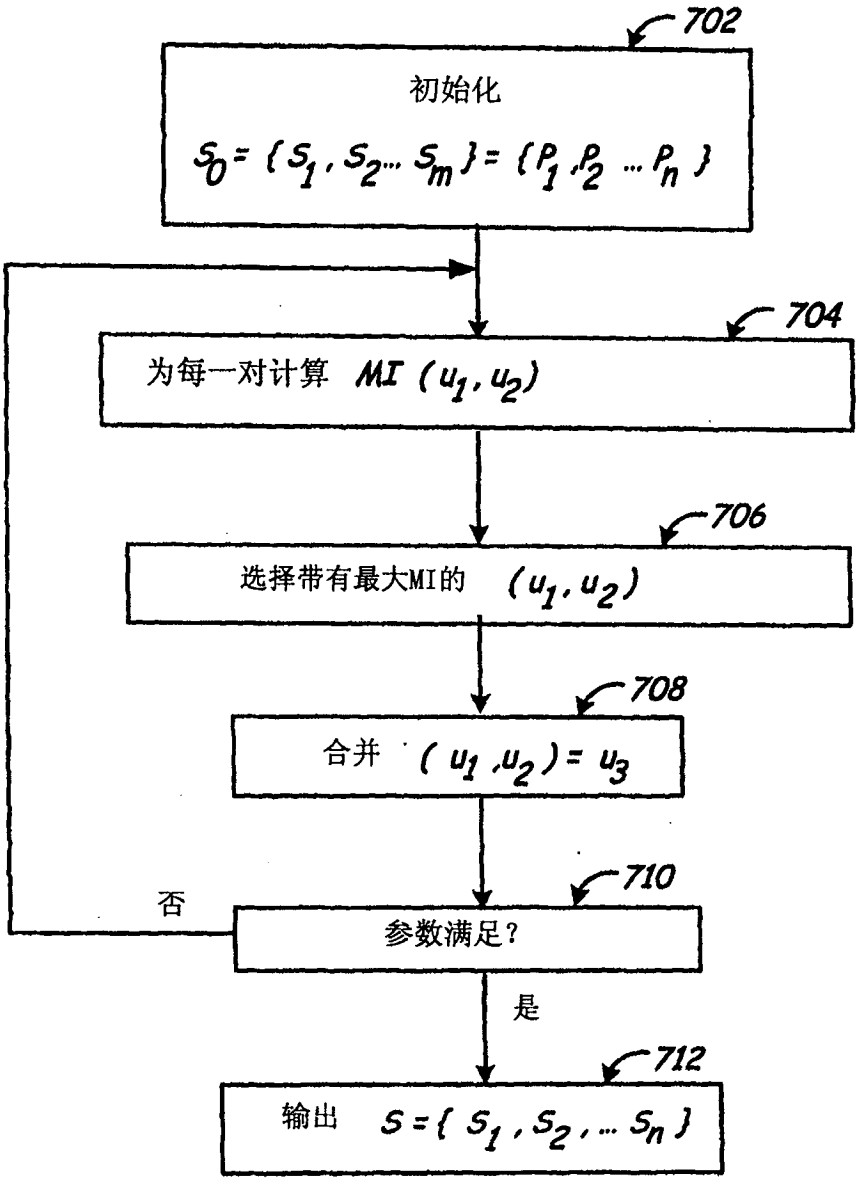

702
初始化
$S_0 = \{S_1, S_2 \ldots S_m\} = \{P_1, P_2 \ldots P_n\}$
704
为每一对计算 $MI(u_1, u_2)$
706
选择带有最大MI的 (u_1, u_2)
708
合并 $(u_1, u_2) = u_3$
710
否
参数满足?
是
712
输出 $S = \{S_1, S_2, \ldots S_n\}$

图 7