(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 102282608 B
(45)授权公告日 2013.06.12

(21)申请号 200980154721.5
(22)申请日 2009.12.03
(30)优先权数据
12/330,921 2008.12.09 US
(85)PCT申请进入国家阶段日
2011.07.15
(86)PCT申请的申请数据
PCT/IB2009/007650 2009.12.03
(87)PCT申请的公布数据
WO2010/067165 EN 2010.06.17
(73)专利权人 诺基亚公司
地址 芬兰埃斯波
(72)发明人 J·田
(74)专利代理机构 北京市金杜律师事务所 11256
代理人 酆迅 姜彦
(51)Int.Cl.
G10L 15/07(2013.01)

(56)对比文件
CN 1223739 A,1999.07.21, 全文.
CN 1783213 A,2006.06.07, 全文.
US 2004230420 A1,2004.11.18, 全文.
US 2005182626 A1,2005.08.18, 全文.
WO 0014729 A2,2000.03.16, 全文.
WO 9622514 A2,1996.07.25, 全文.

审查员 隋欣

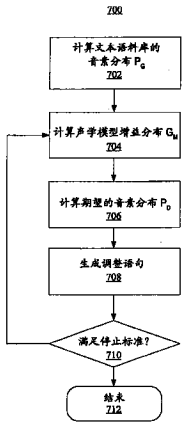
权利要求书2页 说明书10页 附图6页

(54)发明名称

自动语音识别声学模型的调整

(57)摘要

本发明公开了一种用于调整声学模型的方法和系统。用户终端可以确定文本语料库的音素分布；确定调整声学模型之前以及之后、所述声学模型的声学模型增益分布；基于所述音素分布和所述声学模型增益分布来确定期望的音素分布；基于所述期望的音素分布来生成调整语句；以及生成请求用户说出所述调整语句的提示。



CN 102282608 B

1. 一种用于调整自动语音识别声学模型的方法，包括：

确定文本语料库的音素分布；

确定调整声学模型之前以及之后、所述声学模型的音素的声学模型增益分布；

基于所述音素分布和所述声学模型增益分布来确定期望的音素分布；

基于所述期望的音素分布来生成调整语句；以及

生成请求用户说出所述调整语句的提示。

2. 如权利要求 1 的方法，进一步包括：基于根据来自说出所述调整语句的用户语音输入对音素的统计表征进行更新来调整所述声学模型，以生成经更新的声学模型。

3. 如权利要求 2 的方法，进一步包括：基于所述声学模型的音素和所述经更新的声学模型的音素来确定经更新的声学模型增益分布。

4. 如权利要求 3 的方法，进一步包括：确定基于所述经更新的声学模型增益分布的相似性度量满足用以结束调整所述经更新的声学模型的停止标准。

5. 如权利要求 1 的方法，其中所述声学模型增益分布是对调整之前和调整之后的所述声学模型的音素之间相似性进行度量的相似性度量。

6. 如权利要求 1 的方法，其中所述调整语句的生成包括从候选调整语句列表中选择候选调整语句作为所述调整语句。

7. 如权利要求 6 的方法，其中所述调整语句的生成进一步包括：

确定多个候选调整语句的多个候选调整语句音素分布；

标识所述多个候选调整语句中针对所述期望的音素分布具有最小交叉熵度量的第一候选调整语句；以及

将所述第一候选调整语句选作所述调整语句。

8. 如权利要求 1 的方法，其中所述调整语句的生成进一步包括：标识对通过词汇表的多个单词列表分段的累积得分进行优化的单词路径。

9. 如权利要求 8 的方法，其中所述调整语句的生成包括：将词汇表的连续单词列表中单词之间的连接建模为双构词成分，以确定所述单词之间的关系。

10. 如权利要求 1 的方法，其中所述调整语句的生成包括：应用有限状态语法以向所述调整语句提供结构。

11. 如权利要求 1-10 中任一的方法，其中所述音素分布是使用单构词成分语言模型计算的。

12. 一种用于调整自动语音识别声学模型的设备，包括：

用于确定文本语料库的音素分布的装置；

用于确定调整声学模型之前以及之后、所述声学模型的音素的声学模型增益分布的装置；

用于基于所述音素分布和所述声学模型增益分布来确定期望的音素分布的装置；

用于基于所述期望的音素分布来生成调整语句的装置；以及

用于生成请求用户说出所述调整语句的提示的装置。

13. 如权利要求 12 的设备，进一步包括用于基于根据来自说出所述调整语句的用户语音输入对音素的统计表征进行更新来调整所述声学模型以生成经更新的声学模型的装置。

14. 如权利要求 13 的设备，进一步包括用于基于所述声学模型的音素和所述经更新的

声学模型的音素来确定经更新的声学模型增益分布的装置。

15. 如权利要求 14 的设备，进一步包括用于确定基于所述经更新的声学模型增益分布的相似性度量满足用以结束调整所述经更新的声学模型的停止标准的装置。

16. 如权利要求 12 的设备，进一步包括用于从候选调整语句列表中选择候选调整语句作为所述调整语句的装置。

17. 如权利要求 16 的设备，其中所述用于生成所述调整语句的装置进一步包括：

用于确定多个候选调整语句的多个候选调整语句音素分布的子装置；

用于标识所述多个候选调整语句中针对所述期望的音素分布具有最小交叉熵度量的第一候选调整语句的子装置；以及

用于将所述第一候选调整语句选作所述调整语句的子装置。

18. 如权利要求 12 的设备，进一步包括用于标识对通过词汇表的多个单词列表分段的累积得分进行优化的单词路径的装置。

19. 如权利要求 18 的设备，进一步包括用于将词汇表的连续单词列表中单词之间的连接建模为双构词成分以确定所述单词之间的关系的装置。

20. 如权利要求 12-19 中任一的设备，进一步包括用于应用有限状态语法以向所述调整语句提供结构的装置。

自动语音识别声学模型的调整

技术领域

[0001] 本发明的示例性实施方式总体上涉及语音识别。更具体地，本发明的示例性实施方式涉及对声学 (acoustic) 模型进行调整的自动语音识别。

背景技术

[0002] 很多当下的自动语音识别 (ASR) 系统需要用户通过下述操作来显式地训练声学模型：读出预定语句，以便基于该用户的语音特征来调整讲话者无关 (SI) 声学模型，从而改进语音识别性能。

发明内容

[0003] 以下给出了本发明某些示例性实施方式的简单概要，以便提供本发明某些示例性实施方式的基本理解。此概要不是详尽的总览，而且也不意在标识重要元素或关键元素或者勾勒权利要求的范围。以下概要仅以作为以下所提供更详细描述的序言的简单形式给出了某些概念和示例性实施方式。

[0004] 本公开的某些示例性实施方式涉及一种用于调整声学模型的装置、方法和系统。

[0005] 更具体地，根据本公开某些示例性实施方式的方法、装置和系统提供了对声学模型的调整。用户终端可以确定文本语料库 (corpus) 的音素分布、确定调整声学模型之前或之后该声学模型音素的声学模型增益分布、基于该音素分布和该声学模型增益分布来确定期望的音素 (phoneme) 分布、基于期望的音素分布来生成调整语句，以及生成请求用户说出该调整语句的提示。

附图说明

[0006] 通过考虑附图来参考下述描述，可以获得本发明的更完整理解及其优势，在附图中相同的附图标记指示相同的特征，以及其中：

[0007] 图 1 示出了根据本公开示例性实施方式的用户终端。

[0008] 图 2 示出了根据本公开示例性实施方式实现的、用于调整声学模型的用户终端的架构。

[0009] 图 3 示出了根据本公开示例性实施方式的、包括文本语料库、发音词典和语音语料库的训练数据库 (DB)。

[0010] 图 4 示出了根据本公开示例性实施方式的、存储有讲话者无关声学模型和讲话者相关声学模型的声学模型数据库 (DB)。

[0011] 图 5 示出了根据本公开示例性实施方式的、存储有语言模型的语言模型数据库 (DB)。

[0012] 图 6 示出了根据本公开示例性实施方式的、针对词汇表生成的、用于基于统计方法来生成调整语句的单词网格 (word lattice)。

[0013] 图 7 示出了根据本公开示例性实施方式的、用于选择优化调整语句以调整声学模

型的方法。

[0014] 图 8 示出了根据本公开示例性实施方式的、描述了针对不同调整技术的单词识别性能的图表。

具体实施方式

[0015] 在各种实施方式的下述描述中，参考形成本文一部分并且在其中仅通过说明方式示出各种实施方式的附图，其中在这些实施方式中可以实现本发明的一个或多个示例性实施方式。应当理解，在不脱离本发明范围的前提下，可以利用其他实施方式以及做出结构和功能上的修改。

[0016] 图 1 示出了根据本公开示例性实施方式的用户终端。用户终端 102 可以使用声学模型、语言模型和发音词典来执行自动语音识别 (ASR)，以便从人类语音中识别出文本，其中该人类语音经由话音接口输入，以允许用户提供用于控制用户终端 102 操作的语音输入（如下文进一步详述的）。

[0017] 在示例性实施方式中，用户终端 102 可以是所示的移动通信设备、具有天线的移动电话或移动计算机，或者也可以是数字视频记录器 (DVR)、机顶盒 (STB)、服务器计算机、计算机、存储设备、因特网浏览器设备、游戏设备、音频 / 视频播放器、数字相机 / 摄像机、电视、无线电广播接收机、定位设备、有线或无线通信设备和 / 或其任何组合。用户终端 102 可以是如所示的单独设备，或者可以集成在诸如但不限于汽车内的另一设备中。

[0018] 在所示示例中，用户终端 102 包括显示器 104、处理器 106、存储器 108 或其他计算机可读介质和 / 或其他存储、用户接口 110、麦克风 112 和扬声器 114。用户设备 102 的麦克风 112 可以从用户接收语音输入，而扬声器 114 可以输出音频以提示用户与话音接口进行交互。用户接口 110 可以包括小键盘、触摸屏、话音接口、四箭头键、游戏杆、数据手套、鼠标、滚球 (roller ball)、触摸屏或用于从用户接收用以控制用户终端 102 的输入的其他适当设备。

[0019] 图 2 示出了根据本公开示例性实施方式的、用于调整声学模型的用户终端 102 的架构 200。架构 200 的处理器 106 可以通过下述操作来创建讲话者相关模型 :根据从使用有效调整语句的讲话者接收的语音输入来调整讲话者无关模型。架构 200 可以动态地标识用于该调整过程的优化调整语句。

[0020] 在所示示例中，架构 200 可以包括处理器 106，该处理器 106 包括音素分布处理器 204、声学模型增益处理器 206、调整语句处理器 208 和静态音素分布处理器 210。该处理器 106 可以是实现音素分布处理器 204、声学模型增益处理器 206、调整语句处理器 208 和静态音素分布处理器 210 的单个处理器，或者可以是彼此远离或位于彼此本地的两个或更多单独处理器。架构 200 的存储器 108 可以存储数据，其包括语言模型数据库 202、训练数据库 214 和声学模型数据库 216，这些将参考图 3- 图 5 做进一步详细描述。训练数据库 214 还可以是存储器 108 的输入，如所示。

[0021] 图 3 示出了根据本公开示例性实施方式的、包括文本语料库、发音词典和语音语料库的训练数据库。文本语料库 302 可以是包括一种或多种语言文本的结构化集合的数据库。文本语料库 302 可以基于来自于书籍、新闻、单词表、数字序列、多人之间的语音对话等的摘录。发音词典 304 可以包括具有特定发音的单词或短语的集合。在示例性实施方式

中，发音词典可以具有针对文本语料库 302 中每个单词的条目列表，该条目包括单词及其发音。例如，对于单词“you（你）”，发音字典可以列出该单词“you”及其音素级别发音：“ju”。语音语料库 306 可以是包括语音音频文件和每个音频文件的文本转录的数据库。例如，语音语料库 306 可以包括作为说出“How are you doing today？（你今天过得怎么样）”的某人音频记录的音频文件，以及文本转录可以包括对应于该音频记录的文本。

[0022] 图 4 示出了根据本公开示例性实施方式的、存储有讲话者无关声学模型和讲话者相关声学模型的声学模型数据库。如所示，声学模型数据库 216 可以包括一个或多个讲话者无关 (SI) 模型 402 和一个或多个讲话者相关 (SD) 模型 404。SI 模型 402 和 SD 模型 404 二者可以使用预先记录的语音进行训练。在示例性实施方式中，SI 声学模型 402 和 SD 声学模型 404 可以根据训练数据库 214 的文本语料库 302 和语音语料库 306 进行训练。声学模型 402 和 404 可以是例如上下文依赖音素隐形马尔科夫模型 (HMM)。

[0023] 用户终端 102 可以使用声学模型 402 和 404 对接收自特定用户的语音输入进行分类，以便识别出语音输入中说出的单词。声学模型可以包括对不同声音、单词、单词的部分和 / 或其任何组合进行建模的数据，以便在接收自用户的语音输入中识别出单词。声学模型可以包括构成文本语料库 302 中每个单词的声音的统计表征。为了开发出针对多个用户可用的声学模型，声学模型可以根据记录自多个讲话者的语音数据进行训练，并且该声学模型可以称为 SI 声学模型 402。声学模型的训练可以涉及对说出的单词进行统计建模的过程，以使得与该说出的单词对应的文本可以由用户终端 102 识别。

[0024] SI 声学模型 402 例如可以开发自多个个体提供的语音输入，并且由此可以代表一般讲话者的语音特征，但可能未考虑到个体讲话者唯一的讲话特征。该训练过程可以泛化该 SI 声学模型 402，以表征来自特定讲话者的、待识别的说出单词的特征。由于 SI 声学模型 402 开发自多个讲话者，所以 SI 模型 402 可能针对特定讲话者提供的语音不具有较高的单词识别准确度。用户终端 102 可以调整该 SI 声学模型 402 以改进语音识别准确度。此处描述的讲话者调整方案可以利用有限的讲话者特定数据来调整（例如，调谐）该 SI 声学模型 402，以更好地表征该个体讲话者的特征。

[0025] 调整可以从特定讲话者获得有限量的语音输入，以便调整该 SI 声学模型 402 从而生成 SD 声学模型 404。调整可以迭代执行。该 SI 声学模型 402 可以通过记录特定讲话者的语音输入而得以调整，从而生成 SD 声学模型 404 的初始迭代。用户可以提供用以继续调整该 SD 声学模型 404 的进一步输入语音。例如，特定讲话者可以说出用于调整声学模型的一个语句。该特定讲话者可以提供一个或多个附加语句，直到该调整会话完成为止。调整语句的有效设计在下文进一步详细讨论。

[0026] 用户终端 102 可以用作个人设备，诸如但不限于，大部分情况下由单个用户使用的移动电话。在由单个用户使用时，用户终端 102 可以包括适合于该单个用户语音特征的单个 SI 声学模型 404。而且，如果多个用户共享用户终端 102 的话，用户终端 102 可以为每个用户提供多个 SD 声学模型 404。例如，如果用户终端 102 由多个用户共享，该声学模型可以包括适合于每个用户的 SD 声学模型 404。

[0027] 图 5 示出了根据本公开示例性实施方式的、存储有语言模型的语言模型数据库。该语言模型数据库 202 可以存储一个或多个声学语言模型，诸如根据训练数据库 214 的文本语料库 302 和语音语料库 306 训练而来的 502A 和 502B。语言模型 502 可以是向单词序

列指派概率的文件，并且其可以预测语音序列中的下一个单词。在示例性实施方式中，该语言模型（例如，502A 和 502B）可以是 n 构词成分语言模型。n 构词成分语言模型可以是确定观测到具有某个单词序列的语句的概率的模型。例如，单构词成分语言模型可以指示单个单词在文本语料库 302 中出现的频率概率，双构词成分语言模型可以指示双单词序列在文本语料库 302 中出现的频率概率，以及 n 构词成分语言模型可以指示 n 单词序列在文本语料库 302 中出现的频率概率，其中 n 是正整数。在示例性实施方式中，语言模型 502A 可以是单构词成分语言模型，而语言模型 502B 可以是双构词成分语言模型。

[0028] 架构 200 可以解决文本语料库 302 的音素分布问题，以设计出用于高效地调整 SI 声学模型 402 的优化调整语句。语音可以分解成音素，其中音素是子单词单元，但是也可以是其他声学单位。子单词单元的示例是普通话的声韵或者音节。音素的示例是单音或上下文依赖的音素，诸如三音素。音素分布可以度量每个音素出现在文本语料库 302 中的频率。对于具有有限量调整文本的文本语料库 302，某些音素较之于其他音素可能出现得更加频繁。

[0029] 有限量的调整文本可以导致 SI 声学模型 402 针对某些音素具有有限的信息，并且当该用户说出这些音素时，用户终端 102 可能具有较低的单词识别准确度，尤其是在用户的语音特征显著不同于为创建该 SI 声学模型 402 提供语音输入的个体时。而且，读出预定语句可以是非常耗时的任务，通常也不是用户友好的，而且也不能高效地调整该 SI 声学模型 402。为了解决这些和其他问题，架构 200 可以对文本语料库 302 的音素分布进行处理，以高效地创建调整语句，从而实现期望的音素分布，同时将要求用户在有辅导的调整过程期间说出的文本量最小化。

[0030] 在示例性实施方式中，用户终端 102 可以基于用户的语音特征和用户终端 102 在其中进行使用的背景环境来调整该 SI 声学模型 402，以生成 SD 声学模型 404。如下文进一步详细描述的，用户终端 102 可以对该 SI 声学模型 402 进行处理，以生成可以用来调整该 SI 声学模型 402 同时又将来自用户的语音输入量最小化的调整语句。以下描述了这样的方法，其可以使用用于训练语言模型（例如 502A）的文本语料库 302 的音素分布和声学模型增益分布的目标函数（objective function）优化来自动、动态且优化地生成调整语句，以便有效地改进语音识别准确度和用户体验。

[0031] 再次参考图 2，用户终端 102 的架构 200 可以实现这样的自动语音识别（ASR）技术，其可以减轻可能不愿意为了调整 SI 声学模型 402 而执行密集过程的用户的负担。此处讨论的自动语音识别技术可以通过生成用以有效调整 SI 声学模型 402 的优化调整语句而成为较不耗时的任务。

[0032] 用户可以访问用户终端 102 的话音或图形接口，以便开始调整该 SI 声学模型 402。在初次使用该接口期间，用户终端 102 可以执行有辅导的调整过程，在该过程中，该接口请求用户说出预定语句，以便提供用于将 SI 声学模型调整为讲话者相关（SD）声学模型的语音输入。用户终端 102 可以基于该用户的语音特征以及用户终端 102 在其中进行使用的背景环境来调整该 SI 声学模型 402，以开发出 SD 声学模型 404 从而改进单词识别准确度。用于调整 SI 声学模型 402 的语音输入量可以依赖于从训练数据库 214 学习而来的音素分布和用户特质。用户终端 102 例如可以利用有限的用户特定语音输入来调谐 SI 声学模型 402 以调整该声学模型，从而更好地识别出该用户提供的语音。

[0033] 为了开始调整该 SI 声学模型 402，静态音素分布处理器 210 可以确定用于训练语言模型（诸如 502A）的文本语料库 302 的音素分布。音素分布可以代表某些声音在文本语料库 302 中出现的频率。在示例性实施方式中，静态音素分布处理器 210 可以获取语言模型 502A，并且继而基于以下等式来计算文本语料库 302 的静态音素分布 P_G：

[0034] $$P_G = \sum_{i=1}^{V} LM(w_i) \cdot P_w(w), \quad (1)$$

[0035] 其中，LM 可以指示语言模型 502A 可以是单构词成分语言模型，V 可以指示文本语料库 302 的词汇表中不同单词的数量，P_w 可以指示给定的第 i 个单词（其中 i ＝ 1 到 V）的音素分布，以及 w_i 可以指示文本语料库 302 的词汇表中的单词。词汇表可以表示包括在文本语料库 302 中的单词集合。音素分布 P_G 表示为静态的，因为该分布仅依赖于训练数据库 214 的文本语料库 302，并且不可能随时间变化。单词 w_i 的音素分布是音素出现在单词 w_i 中的频率。单构词成分语言模型 LM 502A 可以是单词 w_i 在文本语料库 302 中出现的频率。单构词成分语言模型 LM 502A 可以根据文本语料库 302 训练而来，以及 P_W 可以从具有针对文本语料库 302 词汇表中每个单词 w_i 的音素级别发音的发音词典 304 获得。静态音素分布处理器 210 可以根据等式 (1) 唯一地确定针对给定文本语料库 302 和发音词典 304 的音素分布 P_G。

[0036] 声学模型增益处理器 206 可以针对调整之前和调整之后的声学模型的音素来计算声学模型增益分布 G_M。最初，声学模型增益处理器 206 可以处理 SI 声学模型 402 的音素及其首次调整（即，初始 SD 声学模型 404）。在后续计算中，声学模型增益处理器 206 可以处理 SD 声学模型 404 的不同调整的音素。音素的声学模型增益可以度量针对在调整之前和调整之后的声学模型中定义的每个音素的声学模型增益分布 G_M 的相似性。该调整可以递归地执行。较大的声学模型增益可以指示声学模型（例如，SD 声学模型 404）需要更多数据以用于进一步调整，而较小的声学模型增益可以指示该声学模型接近或者已经达到稳定的已调整状态，不需要更多的调整数据。

[0037] 在示例性实施方式中，声学模型增益处理器 206 可以确定相似性度量 d，以对调整之前和调整之后的声学模型的音素进行比较。针对第 i 个音素，调整之前的声学模型可以是 λ_i，调整之后的声学模型可以是 λ'_i。声学模型增益处理器 206 可以使用每个音素 S 个状态的高斯混合密度模型来计算两个声学模型 λ_i 和 λ'_i 之间的相似性度量 d，其中音素的每个状态 l ＝ 1，2...S 可以利用 N 个高斯概率的混合进行描述。每个高斯混合密度 m 可以具有混合权重 w_m，并且可以具有 L 分量方差 μ_m 和标准差 σ_m。混合权重 w_m 可以是针对每个混合的归一化权重。声学模型增益处理器 206 可以根据下述等式、使用声学相似性度量 d 来计算声学模型增益分布 G_M：

[0038] $$d(\lambda_i, \lambda'_i) = \sum_{l=1}^{S} \sum_{m=1}^{N_{i,l}} w_m^{(i,l)} \cdot \min_{0<n\le N_{i,l}} \sum_{k=1}^{L} \left(\frac{\mu_{m,k}^{(i,l)} - \mu_{n,k}^{'(i,l)}}{\sigma_{n,k}^{(i,l)}} \right)^2 \quad (2)$$

[0039] $$G_M(\lambda_i, \lambda'_i) = \frac{d(\lambda_i, \lambda'_i) + d(\lambda'_i, \lambda_i)}{2} \quad (3)$$

[0040] 其中 i 指示 HMM 的索引，l 指代 HMM 的状态。声学模型增益分布 G_M 可以代表几何混乱度量。声学模型增益分布 G_M 也可以与由声学模型之一在另一个上释放的特征向量的

预期负对数似然度得分的对称近似更加相关，其中混合权重贡献被忽略。

[0041] 音素分布处理器 204 可以基于讲话者无关音素分布 P_G 和声学模型增益分布 G_M 来生成期望的音素分布 P_D，其是讲话者无关的。音素分布处理器 204 继而可以基于下述等式来计算期望的音素分布 P_D：

[0042] $$P_D = \beta \cdot P_G + (1-\beta) \cdot G_M, \qquad (4)$$

[0043] 其中 $0 \leqslant \beta \leqslant 1$ 启发式地设置为控制因子，以平衡讲话者无关音素分布 P_G 和讲话者相关声学模型增益分布 G_M。当 β 接近 1 时，期望的音素分布 P_D 完全依赖于 SI 声学模型 402 的音素分布 P_G，而有辅导的调整过程对于每个讲话者和每个调整过程是相同的。当 β 接近 0 时，期望的音素分布 P_D 完全依赖于声学模型增益分布 G_M，所以有辅导的调整过程对于不同用户乃至同一用户可以不同。因此，β 可以平衡两个分布 P_G 和 G_M，以便更高效地执行。β 可以在制造时进行调谐和预设，或者可以基于用户设置进行调整。调整语句处理器 208 可以使用期望的音素分布 P_D 来生成调整语句。

[0044] 调整语句处理器 208 可以将交叉熵用作目标函数 I，以基于期望的音素分布 P_D 来生成调整语句。交叉熵可以度量似然度比率的预期对数，以检测两个概率分布之间的相似性。调整语句处理器 208 可以通过生成和 / 或选择具有与期望的音素分布 P_D 近似的候选调整语句音素分布 P_n 的并限制了需要用户说出的调整语句量的一个或多个调整语句来优化目标函数 I，由此改善了用户体验。

[0045] 在示例性实施方式中，调整语句处理器 208 可以将交叉熵用作目标函数 I，以度量用于近似期望的音素分布 P_D 的候选调整语句的音素分布 P_n 与期望的音素分布 P_D 之间的音素分布匹配。而且，音素分布 P_n 可以基于多个候选调整语句。期望的音素分布 P_D 可被视为目标分布，而 P_n 可以涉及用于近似目标分布 P_D 的候选调整语句的分布。调整语句处理器 208 可以使用下述等式来计算目标函数 I：

[0046] $$I(P_D, P_n) = \sum_{m=1}^{M} P_{n.m} \cdot \log \frac{P_{n.m}}{P_D} \qquad (5)$$

[0047] 其中 $P_{n.m}$ 是第 n 个候选语句中第 m 个音素的频率，而 M 可以代表音素数量。调整语句处理器 208 可以针对期望的音素分布 P_D 将目标函数 I 最小化，以标识具有离散概率空间中最接近期望的音素分布 P_D 的候选调整语句分布 P_n 的候选调整语句。

[0048] 调整语句处理器 208 可以通过从预定义的候选调整语句列表中选择一个或多个候选调整语句、使用统计方法生成人工调整语句或者使用这些方法的组合来选择候选调整语句。

[0049] 利用语句选择方法，调整语句处理器 208 可以从预定义语句列表中的预定义候选调整语句列表中选择候选调整语句。预定义语句列表可以是由开发者创建的语句列表。该语句选择方法可以选择自然语言语句，但是可能需要具有适度效率的预定义语句列表。自然语言语句可以涉及具有人员在每天的会谈中可能使用的语义含义的语句，这点与可能不具有语义含义的人工生成语句不同。调整语句处理器 208 可以如上所述，通过使用目标函数 I 而从具有大量候选调整语句的文本语料库 302 中选择优化调整语句。更大量的候选调整语句可以用于改进性能，但是可能存在与收集工作量、需要的存储器量以及性能之间的权衡。

[0050] 在示例性实施方式中，调整语句处理器 208 从一个空的候选调整语句集合开始，

可以一次向该语句集合添加一个候选调整语句，直到达到语句要求的数量。语句要求的数量可以依赖于调整效率或者可以设置为常数，诸如但不限于 30 至 50 个语句。当调整产生了 SD 声学模型 404 的标称模型更新时，可以终止调整。调整语句处理器 208 可以从该列表中选择候选调整语句以加入该语句集合，使得具有新添加的候选调整语句的语句集合具有利用上述等式 (5) 的、候选调整语句音素分布 P_n 与期望的音素分布 P_D 之间目标函数 I 的最小交叉熵度量。

[0051] 除了从列表中选择候选调整语句之外，调整语句处理器 208 可以基于相邻单词和 / 或声音之间的统计关系来创建人工调整语句。人工调整语句可以是可能不具有语义含义的单词和 / 或声音的任意集合。调整语句处理器 208 可以优化人工调整语句的设计以改进效率。该设计可以通过缩减开发工作量（因其不需要预先收集调整语句）而得到优化。调整语句处理器 208 可以使用统计方法（如下所述）来生成人工语句。

[0052] 图 6 示出了根据本公开示例性实施方式的、针对词汇表生成的单词网格，用于基于统计方法来生成调整语句。用户终端 102 可以创建具有 n 个单词序列的预定义语句长度的语句，并且单词网格 600 可以是可能单词序列的图形表征。

[0053] 在语句的每个单词分段处，调整语句处理器 208 可以生成单词列表。该单词分段可以是语句中单词的实例，并且该单词列表可以是可能候选单词的列表。调整语句处理器 208 可以确定当前单词列表分段中每个单词与在前单词列表分段中每个单词之间的连接，以标识最佳路径。

[0054] 在示例性实施方式中，调整语句处理器 208 可以使用第一顺序和第二顺序的 n 构词成分，即，单构词成分和双构词成分，以便标识当前单词列表分段 604 中的单词与在前单词列表分段 602 中的单词之间的连接。例如，单词列表可以包括文本语料库 302 中的所有单词。调整语句处理器 208 可以使用双构词成分语言模型 $LM(word_i|word_{i-1}$（单词 $_i$| 单词 $_{i-1}$）) 来对连接进行建模，以便标识在第 (i-1) 个单词列表分段 602 处的单词与第 i 个单词列表分段 604 处的单词之间的连接。双构词成分语言模型可以基于一个单词后面接另一单词的概率来对单词序列进行建模。可以应用令牌通过或 A* 搜索，以找到形成人工调整语句的最佳路径。A* 搜索是公知的最佳优先图形搜索算法，其可以用于找到通过单词网格 600 的最小开销路径。其他方法也可使用。

[0055] 在使用令牌通过时，调整语句处理器 208 可以针对单词网格 600 搜索在第 (i-1) 个单词列表分段 602 处的单词 (word) 与第 i 个单词列表分段 604 处的单词之间的路径（例如，在第 (i-1) 个单词列表分段 602 中的单词 k 与第 i 个单词列表分段 604 中的单词 j 之间的路径），其将在以下等式中提供的累积得分 (accumulative score) 最大化。

[0056]
$$accumulative\_score_i(word_{i,j}) = \max_{k \in Voc} \left\{ \begin{array}{l} accumulative\_score_{i-1}(word_{i-1,k}) + \\ C \cdot LM(word_{i,j} \mid word_{i-1,k}) + \\ I(P_D, P(word_{path}, word_{i,j})) \end{array} \right\} \quad (6)$$

[0057] 在第 i 个单词列表分段 604 处针对第 j 个单词的累积得分 $accumulative\_score_i(word_j)$ 通过查找可以将等式 (6) 中的累积得分最大化的在前第 (i-1) 个单词列表分段 602 的最佳单词 k 来进行更新。C 是声学模型惩罚常量，目标函数 I 是从语句中的第一个单词到当前单词的实际音素分布与期望的音素分布 P_D 之间的交叉熵度量。声学模型惩罚常量 C 可以平衡来自语言模型和来自声学模型的贡献。

[0058] 在达到预定义的语句长度时，调整语句处理器 208 可以基于最终累积得分来对各个单词列表分段中的单词之间的路径进行排名。调整语句处理器 208 可以将具有最高累计得分的路径选作调整语句。调整语句处理器 208 可以生成要求用户说出调整语句的提示，以提供用于调整 SD 声学模型 404 的语音输入，从而通过基于用户说出的调整语句对音素的统计表征进行更新来生成 SD 声学模型 404 的更新。

[0059] 上述人工调整语句方法可以有效地生成优化调整语句，但是该优化调整语句可能是单词的无意义聚集，因为创建该调整语句是为了提供期望的声音聚集而不是为了提供语义含义。由于调整语句可以用来调整声学模型，所以语句的语义含义并不总是那么重要。然而，调整语句处理器 208 可以实现句法结构，以便提供具有合理语义含义的、生成的人工调整语句。为了改进调整语句的语义含义，调整语句处理器 208 可以使用有限状态语法 (FSG) 和基于类 (class) 的语言模型。FSG 可以代表语言模型中的多个类结构。调整语句处理器 208 可以使用 FSG 的结构来为生成的调整语句提供结构，使得该人工调整语句提供完成该语言模型中的类的单词。例如，基于类的语言模型可以是 ：

[0060] $Person_Name_Class（人名类）；例如，John，Smith，

[0061] $Location_Name_Class（地点名称类）；例如，Boston，Paris

[0062] $Natural_Number_Class（自然数类）；例如，21

[0063] $Digit_Class（数字类）；例如，21，

[0064] $Time_Class（时间类）；例如，2:30，

[0065] $Date_Class（日期类）；例如，2008 年 7 月 30 日。

[0066] FSG 可以是 ：

[0067] 语句开始 $Person_Name_Class 预定在 $Date_Class、$Time_Class 从 $Location_Name_Class 到 $Location_Name_Class 的航班语句结束。

[0068] 调整语句处理器 208 可以通过使用上述方法标识出用以使用 FSG 的结构完成人工调整语句以生成该人工调整语句（例如，John 预定在 2008 年 7 月 30 日 2:30 分、从 Boston 到 Helsinki 的航班）。由于 FSG 的结构约束，调整语句处理器 208 可以生成具有语义含义的人工调整语句。一旦已经标识出优化调整语句，无论是通过从列表中选择还是通过人工创建，调整语句处理器 208 都可以生成请求用户说出调整语句的提示，以提供用于调整 SD 声学模型 404 的语音输入，从而通过基于用户说出的调整语句对音素的统计表征进行更新来生成 SD 声学模型 404 的更新。

[0069] 在已使用调整语句调整了 SD 声学模型 404 之后，声学模型增益处理器 206 可以使用上述等式 (2) 和 (3) 来确定用于经更新的 SD 声学模型 404 的相似性度量 d，以生成声学模型增益分布 G_M 的更新。声学模型增益处理器 206 继而可以使用经更新的声学模型增益分布 G_M 来确定是否进一步调整期望的音素分布 P_D。例如，较大的声学模型增益分布 G_M 可以指示 SD 声学模型 404 需要进一步调整，而较小的声学模型增益分布 G_M 可以指示该 SD 声学模型 404 接近或者已经达到稳定的已调整状态，不需要更多的调整。

[0070] 如果声学模型增益分布 G_M 足够小，则声学模型增益处理器 206 可以确定不进一步调整该 SD 声学模型 404。用户终端 102 的话音接口可以输出音频以通知用户有辅导的调整过程已经完成。

[0071] 如果声学模型增益分布 G_M 不够小，则声学模型增益处理器 206 可以确定进一步调

整该 SD 声学模型 404。音素分布处理器 204 可以利用经更新的声学模型增益分布 G_M 和音素分布 P_G、使用上述等式 (4) 来更新期望的音素分布 P_D。音素分布处理器 204 可以向调整语句处理器 208 传达经更新的期望的音素分布 P_D，以使用上述候选调整语句选择方法和 / 或人工调整语句生成方法来设计另一调整语句。可以继续更新声学模型，直到声学模型增益分布 G_M 足够小。

[0072] 图 7 示出了根据本公开示例性实施方式的、用于选择优化调整语句以调整声学模型的方法。该方法可以从框 702 开始。

[0073] 在框 702 中，用户终端 102 的静态音素分布处理器 210 可以计算音素分布 P_G。静态音素分布处理器 210 可以确定用于训练语言模型 502A 的文本语料库 302 的音素分布。音素分布可以代表某些声音在用于训练语言模型 502A 的文本语料库 302 中出现的频率。在示例性实施方式中，静态音素分布处理器 210 可以获取语言模型 502A，并继而根据等式 (1) 来计算发音词典 304 和文本语料库 302 的音素分布 P_G。

[0074] 在框 704 中，声学模型增益处理器 206 可以计算调整之前和之后的声学模型音素的声学模型增益分布 G_M。在初次通过框 704 时，声学模型增益处理器 206 可以确定 SI 声学模型 402 的音素的声学模型增益分布 G_M 及其首次调整 (即，初始 SD 声学模型 404)，并且在后续计算中，声学模型增益处理器 206 可以使用上述等式 (2) 和 (3) 来处理 SD 声学模型 404 的不同调整的音素。

[0075] 在框 706 中，音素分布处理器 204 可以计算期望的音素分布 P_D。音素分布处理器 204 可以组合声学模型增益分布 G_M 和音素分布 P_G，以使用上述等式 (4) 来确定期望的音素分布 P_D。

[0076] 在框 708 中，调整语句处理器 208 可以基于期望的音素分布 P_D 来生成调整语句。调整语句处理器 208 可以选择具有与期望的音素分布 P_D 最匹配的音素分布 P_n 的调整语句。在示例性实施方式中，调整语句处理器 208 可以确定预定义语句列表中多个候选调整语句的候选调整语句音素分布 P_n，并且可以基于等式 (5) 的目标函数 I 来标识候选调整语句中针对期望的音素分布 P_D 具有最小交叉熵度量的候选调整语句 (即，该候选调整语句具有与期望的音素分布 P_D 最接近的音素分布 P_n)。而且，调整语句处理器 208 可以自动地使用上述方法来生成人工调整语句。用户终端 102 继而可以生成要求用户说出调整语句的提示，以通过基于说出调整语句的用户的语音输入来对 SD 声学模型 404 的音素统计表征进行更新而调整该 SD 声学模型 404。

[0077] 在框 710 中，声学模型增益处理器 206 可以确定是否满足了停止标准。该停止标准可以如上所述基于声学模型增益分布 G_M 的值。如果未满足停止标准，则方法 700 可以返回框 704，以进一步调整该声学模型。如果已满足停止标准，则该方法可以继续到框 712 并结束。

[0078] 图 8 示出了根据本公开示例性实施方式的、描述了针对不同调整技术的单词识别性能的图表。图表 800 示出了针对不同调整技术调整量与时间之间的关系，以描绘识别准确度如何随着时间变化。如所示，存在代表识别准确度的不同线 802-808，其中线 806 代表没有调整，线 808 代表存在离线调整，线 802 代表存在离线和在线调整二者，以及线 804 代表存在在线调整而不存在离线调整。离线调整涉及上述有辅导的调整过程。在线调整表示随着用户终端 102 基于在用户使用话音接口时接收的反馈而随时间调整 SD 声学模型 404

时的调整过程。例如，假定存在用户语音，用户终端 102 可以将该语音解码成文本，并使用识别出的文本来进一步调整 SD 声学模型 404。在此示例中，调整语句处理器 208 可以使用声学贝叶斯调整。在试验中使用的文本集合包括来自 23 位美式英语讲话者（有男性也有女性）的总计 5500 条短消息服务 (SMS) 消息，其中每个讲话者提供 240 个发言。在有辅导的调整期间，声学模型要求每个人说出 30 个登记发言。

[0079] 如图 8 所示，离线有辅导调整（参见线 808）提供了显著的改进，这要归因于可靠的有辅导数据以及发音上较丰富的转录。结合的离线有辅导和在线无辅导调整（参见线 802）带来了最佳性能。由此，有辅导的调整尤其在初次使用话音接口期间将带来最好的识别性能。

[0080] 此处描述的自动语音识别 (ASR) 技术可以克服设备具有受限接口（诸如在移动环境中）的挑战。自动语音识别 (ASR) 技术可以提供改进的用户接口，尤其对于具有有限键盘的移动设备而言。

[0081] 上述 ASR 技术可以用来利用新的调整语句来替换 SI 声学模型中预先选择的调整语句，从而利用来自用户的较少语音输入来调整 SI 声学模型。例如，SI 声学模型可以含有具有不平衡音素分布的预先选择的调整语句，并且由此，使用预先选择的调整语句可能不能有效地调整声学模型。因此，对于声学模型的有辅导讲话者调整，上述 ASR 技术可以高效地设计优化调整语句，以提供最佳单词识别性能，同时将需要用户说出以调整该声学模型的文本量最小化。

[0082] 用户终端内的处理器 106 和其他部件使用的计算机可执行指令和数据可以存储在存储器 108 中，以便执行此处描述的任一方法步骤和功能。存储器 108 可以利用只读存储器模块或随机访问存储器的任何组合实现，可选地包括易失性存储器和非易失性存储器。而且，用户终端 102 计算机可执行指令中的部分或全部可以具体化在硬件或固件（未示出）中。

[0083] 尽管在图 1 中仅描述了每个设备的单个实例，但是用户终端 102 可以包括这些设备中每个设备的一个或多个。而且，图 1 所示每个设备执行的功能可以分布在附加的设备中，或者示出的设备可以彼此组合。此外，用户终端 102 也可以包括在其他系统（未示出）中，或者可以包括附加的设备。例如，用户设备 102 可以集成至汽车中。

[0084] 针对调整声学模型而提供的前述描述提供了具有改进的识别准确度的话音接口。应当理解，此处描述的原理可以扩展至其他自动语音识别技术。而且，以上描述在各种示例性实施方式中描绘了由某些设备执行的某些部件和功能。各种示例性实施方式的部件和功能可以彼此组合和 / 或分离。

[0085] 尽管以特定于结构特征和 / 或方法动作的语言描述了主题，但是应当理解，在所附权利要求中定义的主题不需受限于上述特定特征或动作。相反，上述特定特征和动作仅作为实现权利要求的示例性形式公开。

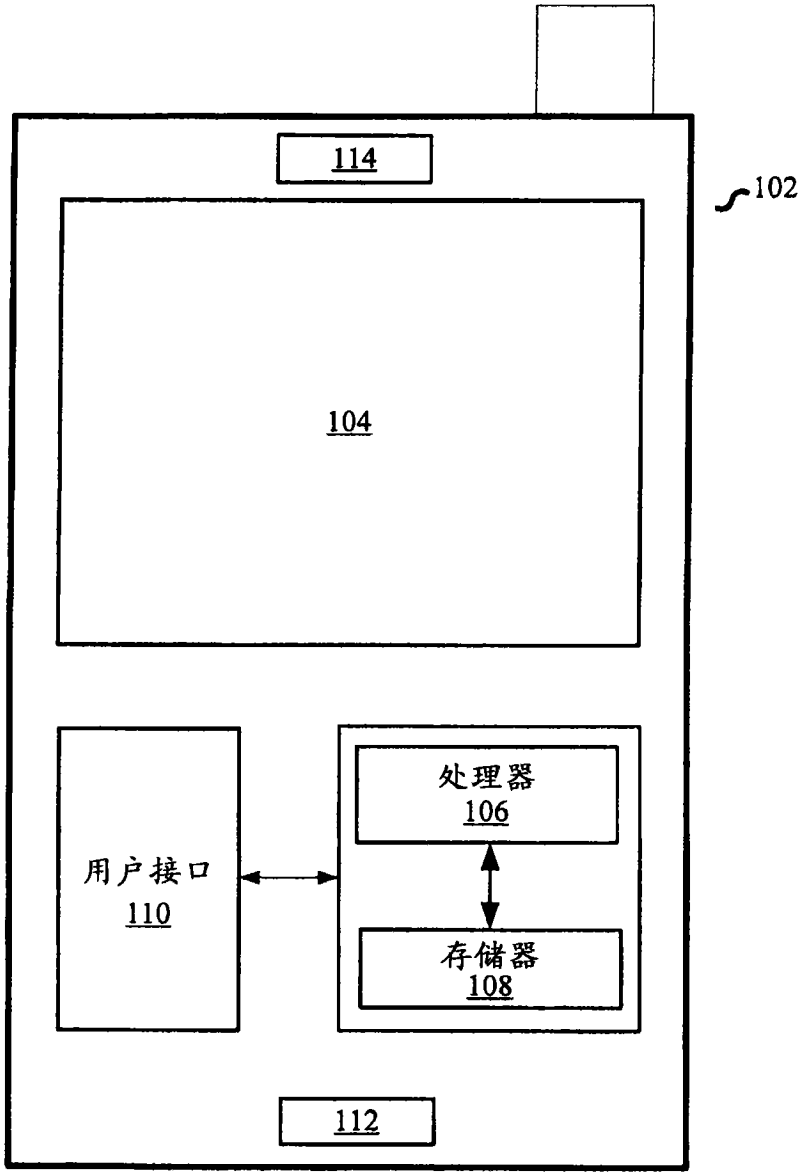


图 1

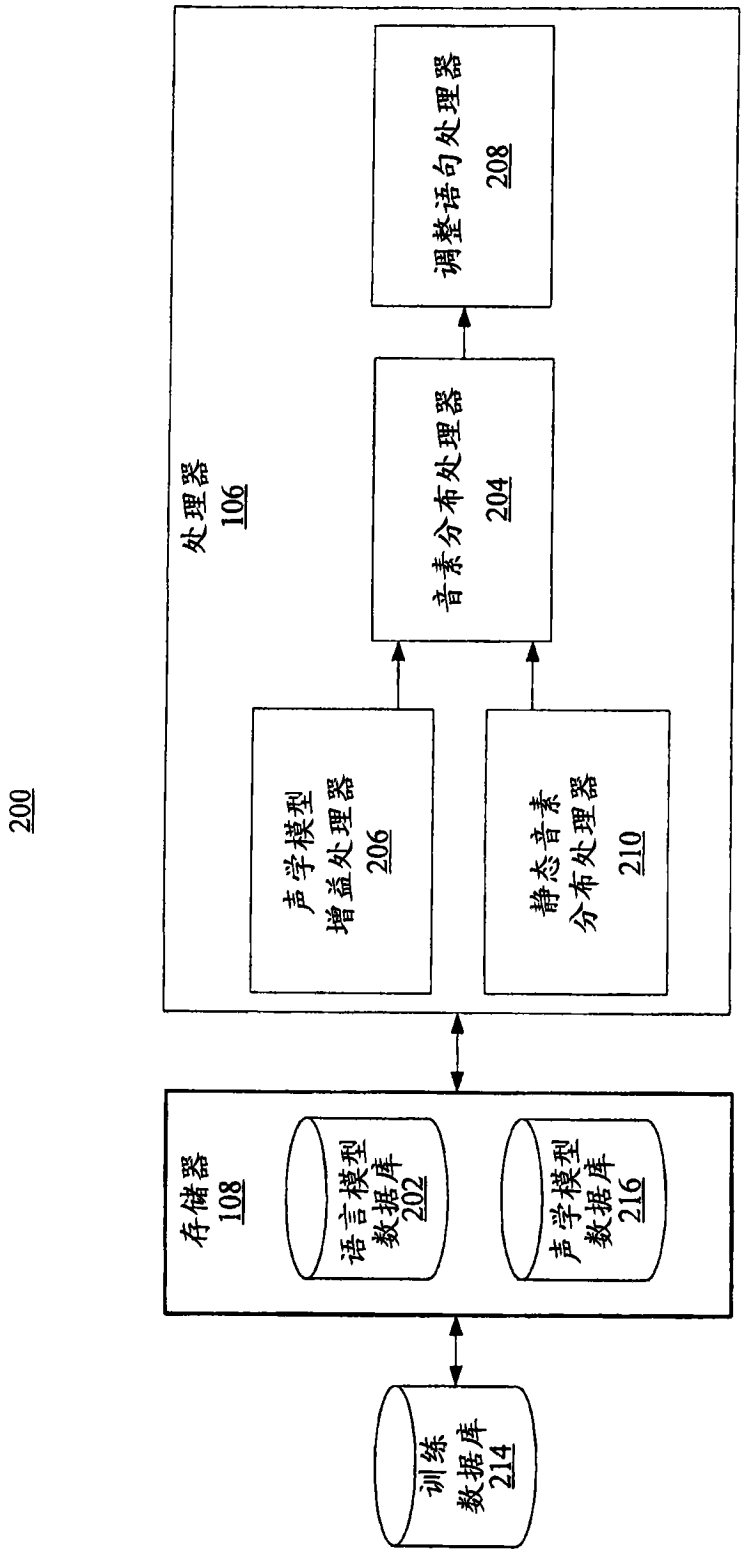

200
处理器
106
调整语句处理器
208
音素分布处理器
204
声学模型
增益处理器
206
静态音素
分布处理器
210
存储器
108
语言模型
数据库
202
声学模型
数据库
216
训练
数据库
214

图 2

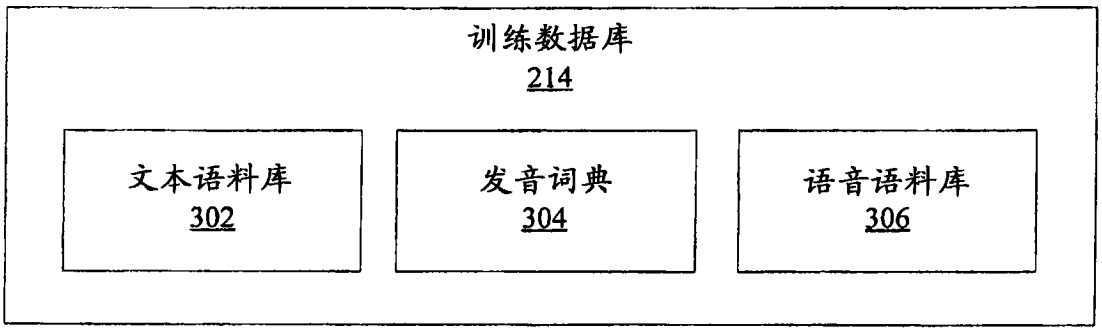

训练数据库
214
文本语料库
302
发音词典
304
语音语料库
306

图 3

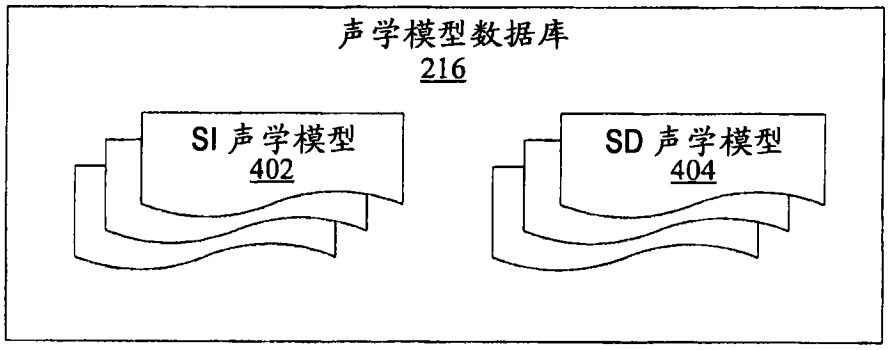

声学模型数据库
216
SI 声学模型
402
SD 声学模型
404

图 4

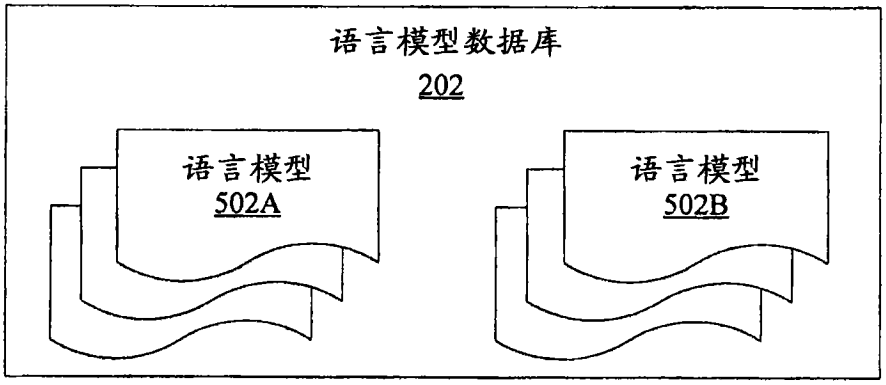

语言模型数据库
202
语言模型
502A
语言模型
502B

图 5

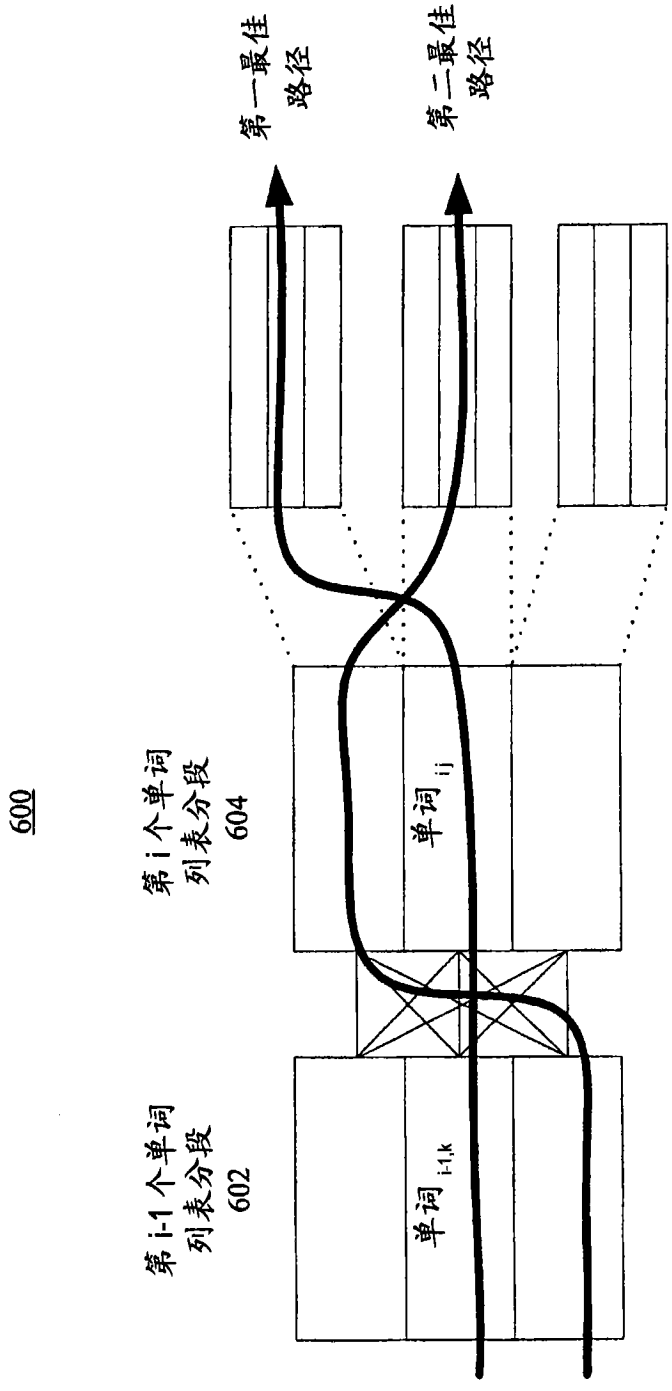


图 6

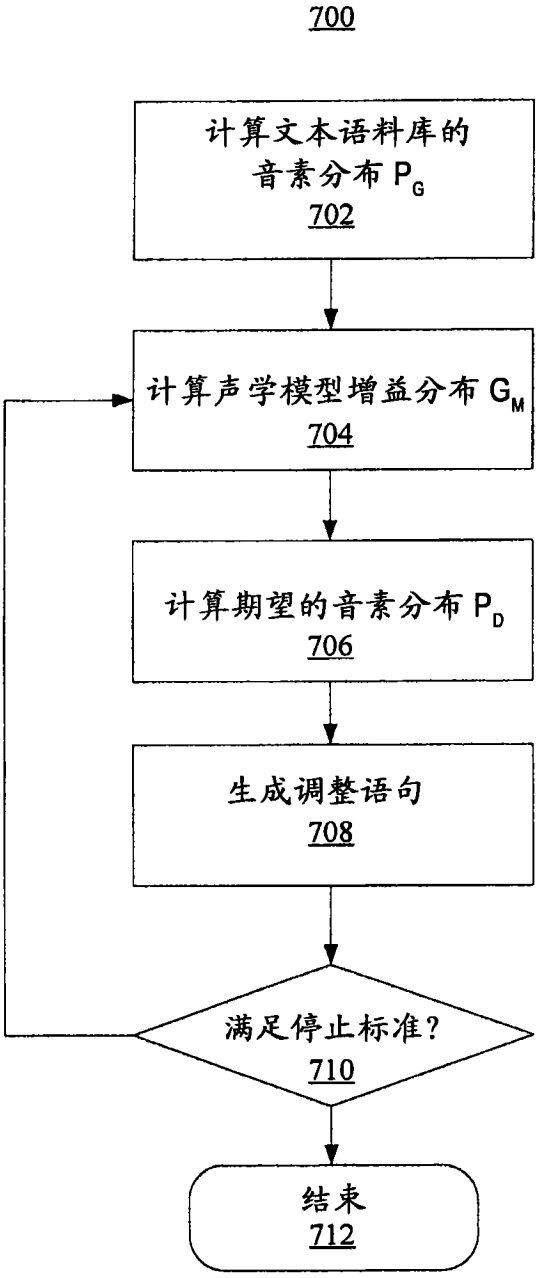

700
计算文本语料库的音素分布 P_G
702
计算声学模型增益分布 G_M
704
计算期望的音素分布 P_D
706
生成调整语句
708
满足停止标准?
710
结束
712

图 7

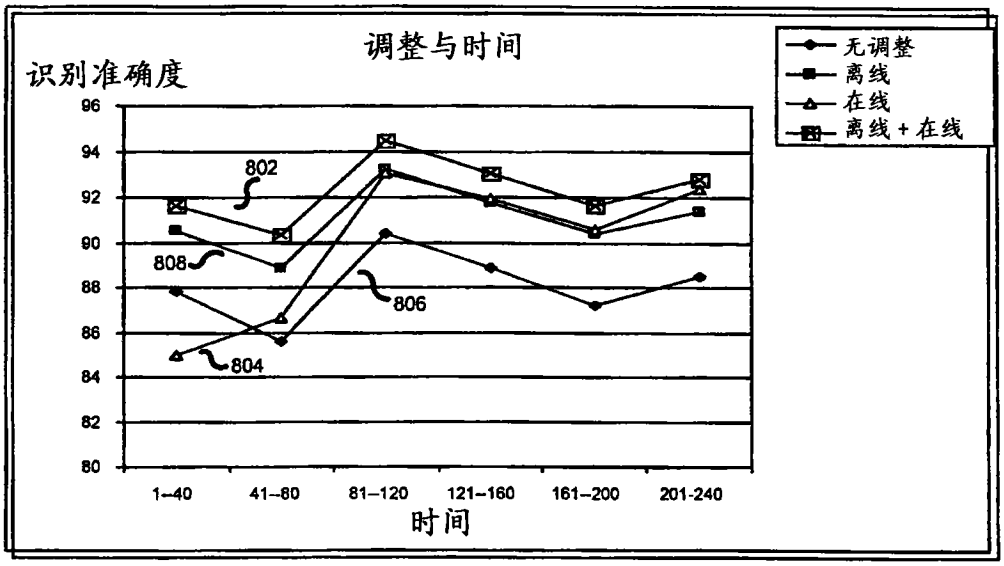

调整与时间
识别准确度
96
94
92
90
88
86
84
82
80
802
808
806
804
1–40
41–80
81–120
121–160
161–200
201-240
时间
无调整
离线
在线
离线 + 在线

图 8