(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105122354 A
(43)申请公布日 2015.12.02

(21)申请号 201380064858.8
(22)申请日 2013.12.10
(30)优先权数据
13/712,891 2012.12.12 US
(85)PCT国际申请进入国家阶段日
2015.06.11
(86)PCT国际申请的申请数据
PCT/US2013/074192 2013.12.10
(87)PCT国际申请的公布数据
WO2014/093384 EN 2014.06.19
(71)申请人 亚马逊技术有限公司
地址 美国内华达州
(72)发明人 伯乔恩·霍夫迈斯特
休·埃文·塞克-瓦尔克
杰弗瑞·科尔内留斯·奥尼尔
(74)专利代理机构 北京天昊联合知识产权代理有限公司 11112
代理人 顾丽波 李荣胜
(51)Int.Cl.
G10L 15/32(2006.01)
G10L 15/30(2006.01)

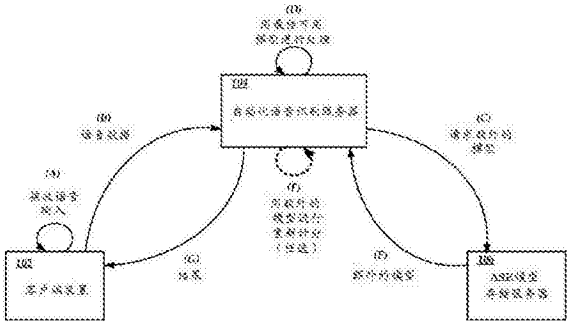
权利要求书2页 说明书14页 附图6页

(54)发明名称

分布式语音识别系统中的语音模型检索

(57)摘要

本发明公开用于管理自动化语音识别系统中的语音识别模型和数据的使用的特征。在被接收到的时候或在话语开始用更通用或不同的模型进行处理之后，模型和数据可被异步检索和使用。一旦被接收到，所述模型和统计数据便可被高速缓存。更新模型和数据所需的统计数据也可被异步检索，从而可以在可用的时候用来更新所述模型和数据。所述经更新的模型和数据可以立即用来再处理话语，或经保存用于处理随后接收的话语。可跟踪用户与所述自动化语音识别系统的交互，以便预测用户可能会使用所述系统的时间。模型和数据可以基于此类预测进行预先高速缓存。



CN 105122354 A

1. 一种系统，其包括：

存储可执行指令的计算机可读存储器；以及

与所述计算机可读存储器通信的一个或多个处理器，其中所述一个或多个处理器经所述可执行指令编程以：

从客户端装置接收包括用户话语的音频数据；

确定额外语音识别模型不可用；

使用基础语音识别模型来对所述音频数据执行第一语音识别处理，以产生第一语音识别结果；

从网络可访问的数据存储区请求所述额外语音识别模型，其中所述请求是在完成所述第一语音识别处理之前开始的；

从所述网络可访问的数据存储区接收所述额外语音识别模型；

使用所述额外语音识别模型以及使用所述音频数据或所述语音识别结果中的至少一个来执行第二语音识别处理；以及

至少部分基于所述第二语音识别处理，将响应传输到所述客户端装置。

2. 根据权利要求 1 所述的系统，其中所述基础语音识别模型包括通用声学模型、性别特定声学模型或通用语言模型中的至少一个，并且其中至少部分基于与所述用户话语相关联的用户的特性来选择所述额外语音识别模型。

3. 根据权利要求 1 所述的系统，其中所述一个或多个处理器还经所述可执行指令编程以：

从所述客户端装置接收包括第二用户话语的第二音频数据；

确定所述额外语音识别模型可用；以及

使用所述额外语音识别模型对所述第二音频数据执行语音识别处理。

4. 根据权利要求 1 所述的系统，其中所述一个或多个处理器还经所述可执行指令编程，以使用多线程处理以与所述第一语音识别处理的执行并行检索所述额外语音识别模型。

5. 根据权利要求 1 所述的系统，其中所述一个或多个处理器还经所述可执行指令编程以高速缓存所述额外语音识别模型。

6. 一种计算机实施的方法，其包括：

在以特定计算机可执行指令配置的一个或多个计算装置的控制下，

对关于用户话语的音频数据执行第一语音处理，以产生语音处理结果；

从网络可访问的数据存储区请求语音处理数据，其中所述请求是在完成所述第一语音处理之前开始的；

从所述网络可访问的数据存储区接收所述语音处理数据；以及

使用所述语音处理数据以及所述音频数据或所述语音处理结果中的至少一个来执行第二语音处理。

7. 根据权利要求 6 所述的计算机实施的方法，其还包括：

至少部分基于所述用户的特性来选择待请求的语音处理数据。

8. 根据权利要求 7 所述的计算机实施的方法，其中所述用户的所述特性包括所述用户的性别、年龄、地域口音或身份。

9. 根据权利要求 6 所述的计算机实施的方法，其中所述语音处理数据包括以下至少一个：声学模型、语言模型、语言模型统计数据、约束最大似然线性回归（“CMLLR”）变换、声道长度归一化（“VTLN”）扭曲因子、倒谱均值和方差数据、意图模型、所命名实体模型或地名录。

10. 根据权利要求 9 所述的计算机实施的方法，其还包括：

请求用于更新所述语音处理数据的统计数据，其中针对统计数据的所述请求是在完成所述第一语音处理之前开始的。

11. 根据权利要求 10 所述的计算机实施的方法，其还包括至少部分基于所述统计数据和所述第二语音处理的结果来更新所述语音处理数据。

12. 根据权利要求 6 所述的计算机实施的方法，其还包括高速缓存所述语音处理数据。

13. 根据权利要求 12 所述的计算机实施的方法，其还包括：

接收关于所述用户的第二话语的第二音频数据；

从高速缓存检索所述语音处理数据；以及

使用所述语音处理数据对所述第二音频数据执行语音处理。

14. 根据权利要求 12 所述的计算机实施的方法，其中高速缓存所述语音识别数据包括：

确定所述用户可能会开始语音识别会话的时间；以及

大体在所述所确定的时间高速缓存所述语音识别数据。

15. 根据权利要求 6 所述的计算机实施的方法，其还包括：

从所述用户操作的客户端装置接收所述音频数据；以及

将响应传输到所述客户端装置，所述响应至少部分基于所述第二语音识别处理。

分布式语音识别系统中的语音模型检索

背景技术

[0001] 现代语音识别系统通常包含声学模型和语言模型。声学模型用来生成关于哪些字词或子字单元（例如，音素）基于话语的声学特征对应于话语的假设。语言模型用来基于说出话语的语言的词汇特征来确定使用声学模型生成的哪个假设最有可能是话语的转录。

[0002] 语音识别中使用的声学模型、语言模型及其它模型（统称为语音识别模型）可在各种程度上专门化或自定义。例如，语音识别系统可具有并不采用任何特定方式自定义的通用模型或基础模型，以及用于特定性别、年龄范围、地域口音或其任何组合的任何数量的额外模型。一些系统可具有用于特定主题（例如，医学术语）乃至特定用户的模型。

[0003] 语音识别系统可以基于客户端或基于客户端－服务器。例如，膝上型计算机等计算装置可包含应用软件和数据，以便将音频输入处理成文本输出或音频输入的可能转录的列表。一些语音识别通过个人或移动计算装置来接受音频输入，并将音频输入传递到网络可访问的服务器，在该网络可访问的服务器中，音频输入被转录或执行其它处理。

附图说明

[0004] 现在将参考以下图式来描述各种发明特征的实施例。贯穿附图中，参考编号可再用来表示所参考元件之间的对应关系。提供图式是为了说明本文中描述的示例性实施例，而不意图限制本发明的范围。

[0005] 图 1 为其中可实施分布式语音识别系统的说明性网络环境的框图，示出了客户端装置、语音识别服务器与模型存储服务器之间的说明性交互。

[0006] 图 2 为说明性语音识别服务器的框图，示出了各种模型和数据存储区。

[0007] 图 3 为用于管理分布式语音识别系统中的语音识别会话的说明性过程的流程图。

[0008] 图 4 为用于在分布式语音识别系统中利用模型的预先高速缓存的说明性过程的流程图。

[0009] 图 5A 和图 5B 为客户端装置、语音识别服务器、模型高速缓存与模型存储服务器之间的说明性交互的框图。

具体实施方式

[0010] 前言

[0011] 大体而言，本发明涉及管理分布式语音识别系统的操作，所述分布式语音识别系统包含专用或自定义语言模型、专用或自定义声学模型以及其它数据，统称为语音识别模型。语音识别系统使用语音识别模型将用户的话语处理成话语的转录或可能转录的列表。一些语音识别系统使用适用于大量用户的通用或基础语音识别模型。在一些情况下，对于个别用户或一组用户而言，语音识别系统可使用额外的模型来提供比基础模型更准确的结果。此类额外的模型可包含或强调特定用户通常使用的词汇，或者其可能与语音识别处理期间以数字方法表示特定用户的语音的方式更紧密匹配。然而，额外的模型（以及一般而言，语音识别模型）可消耗大量的存储空间，因此，语音识别系统在可本地存储在进行语音

识别处理的装置上的模型数量方面受到限制。此外，由于尺寸较大，因此，从其它装置（例如，存储服务器）中检索额外的模型可能会不利地影响用户感知的性能。例如，从存储服务器中检索较大额外模型所需的时间会增加用户在说出话语与接收结果之间经历的延迟。

[0012] 本发明的方面涉及用于对话语执行语音识别的额外语音识别模型的异步检索。在开始处理话语之前或与此并行，语音识别服务器或引擎可从数据存储区请求语音识别模型，从而使得语音识别模型的检索不会干扰初始处理。例如，在多线程系统中，语音识别模型的线程管理检索并不妨碍处理线程。

[0013] 在一些实施例中，语音识别系统可实施为分布式系统，其包含用于执行语音识别的部件（例如，语音识别服务器）和用于存储额外语音识别模型的部件（例如，长期存储服务器）。语音识别服务器可接收来自用户的音频输入，并且从存储部件中检索不同程度自定义或专门化的一个或多个语音识别模型（例如，一个用于用户的性别、一个用于用户的地域口音、一个用于特定用户等）。语音识别服务器可检索额外的语音识别模型，同时还用基础语音识别模型来处理所接收的音频输入。在一些情况下，当请求额外的模型时，可能会有延迟，直到通过网络接收到所述模型为止。这可导致响应于用户话语提供转录或执行动作时出现延迟。用户可能认为这种延迟是无法接受的。然而，如果能足够快地接收到可以使用的额外模型，同时仍为用户提供满意的性能（例如，延迟 <100ms、<500ms 等），那么额外的模型可用来提高语音识别的准确性。例如，在用基础模型开始处理音频输入之前，可接收到额外的语音识别模型，且在这种情况下，从一开始就可使用额外的语音识别模型。作为另一实例，所述模型可能会在用基础模型处理音频输入的过程中或在处理已经完成之后到达。额外的模型可用来再处理音频输入或初始处理的结果，前提是这种再处理可以足够快地完成，以向用户提供满意的性能。

[0014] 除了在处理之前或处理过程中请求额外的语音识别模型之外，语音识别服务器还可异步请求统计数据和其它数据，以更新额外的语音识别模型。额外的语音识别模型可在语音识别服务器处理话语之后被更新。用来更新额外语音识别模型的数据量通常显著大于额外语音识别模型本身中的数据量。有利的是，通过异步请求统计数据和其它数据来更新额外语音识别模型，所述额外语音识别模型可在统计数据和其它数据一旦被接收后就更新。经更新的语音识别模型随后可再次用来提供更准确或在其它方面更好的结果。例如，经更新的语音识别模型可用来再处理更新所依据的当前话语，或者经更新的语音识别模型可用来处理随后的话语，或进行这两者。

[0015] 本发明的另外方面涉及高速缓存额外的语音识别模型。通过高速缓存额外的语音识别模型，它们可被立即使用或大体更快地使用，以用于处理随后接收的话语，从而在与使用基础语音识别模型处理话语大体相同的时间量内提供更准确的结果。例如，语音识别服务器可检索额外的语音识别模型来处理关于从客户端装置接收的话语的音频数据。不论额外的语音识别模型是否在将要处理第一话语的时间到达，它们都可被高速缓存并用来处理关于第二话语的随后接收的音频数据。

[0016] 本发明的其它方面涉及基于对可请求哪些额外模型以及可请求额外模型的时间的预测，预先高速缓存额外的语音识别模型。例如，可监视用户与语音识别系统的交互，从而语音识别系统的部件可检测用户可能使用语音识别系统的模式，或者预测用户将来可能使用语音识别系统的时间。在预期此类使用的情况下，可能将被请求的额外语音识别模型

可被预先高速缓存（例如，从长期存储中检索并存储在语音识别服务器或某一网络可访问的高速缓存部件上）。

[0017] 尽管出于说明的目的，本发明所描述的实施例的各方面将着重于语音识别服务器接收关于话语的音频数据，以及异步检索额外的语音识别模型来处理音频数据，但所属领域的技术人员将了解，本文中公开的技术可应用于任何数量的软件处理或应用。例如，用户的个人移动装置可包含语音识别引擎，并且在话语的本地处理过程中，异步请求待使用的额外的语音识别模型。现在将相对于某些实例和实施例来描述本发明的各方面，这些实例和实施例意图说明而非限制本发明。

[0018] 参考说明性实例，用户可发出声音命令或以其它方式口头上与客户端装置（例如，移动电话或平板计算机）交互。客户端装置可将关于用户话语的数据传输到网络可访问的语音识别服务器，所述语音识别服务器作为分布式自动化语音识别（“分布式 ASR”）系统的一部分。语音识别服务器可使用各种类型的语音识别模型（例如，声学模型和语言模型），以处理话语并且转录或以其它方式确定用户说了什么。为了提高准确性，模型可在各个层次为用户自定义。语音识别服务器可使用基础模型、用于性别、年龄、地域口音、术语等的模型。语音识别模型还可针对特定用户或针对特定时间、日期等自定义（例如，用于假日术语的语言模型）。额外的语音识别模型可能比较大，因此，语音识别服务器可能没有足够的存储容量来存储每个额外的模型。利用额外语音识别模型的分布式 ASR 系统可针对额外模型实施长期存储，从而使得语音识别引擎可使用的每个额外语音识别模型均可被存储并根据需要提供到语音识别引擎。

[0019] 分布式 ASR 系统的用户体验可在质量（例如，结果的准确性）和所感知性能（例如，说出话语与接收到结果之间的等待时间和逝去的时间）两个方面进来定义。分布式 ASR 系统努力尽快返回结果。然而，分布式和其它网络系统固有的等待时间会直接影响用户体验。因此，由于从长期存储中检索额外的语音识别模型而造成的任何额外延迟都可能导致并非令人满意的用户体验。

[0020] 为了最小化使用额外的语音识别模型可能对分布式 ASR 系统带来的负面影响，可异步请求额外的模型（例如，额外语音识别模型的检索不会妨碍用其它模型来执行语音识别过程，且反之亦然）。例如，语音识别服务器可利用多线程处理来请求额外的模型，并且以并行或异步的方式用基础模型来执行语音识别。当接收到话语或关于话语的数据时，语音识别服务器可确定说话人的身份和 / 或说话人的特性（例如，性别）。在处理话语之前、并行或之后，语音识别服务器可检索额外的语音识别模型。由于检索不同种类的额外语音识别模型可能具有不同的等待时间，因此，语音识别服务器或分布式 ASR 系统的某一其它部件可请求任何数量的不同额外模型，并且使用在将要使用模型的时间接收到的一个最好的模型且在不会不利影响用户体验的情况下返回结果。例如，语音识别服务器可请求用于个别用户的模型，且还请求用于用户性别的模型。如果用于性别的模型首先被接收到，那么语音识别服务器可继续使用性别特定的额外语音识别模型来处理话语。然而，如果在将要使用用于特定用户的模型的时间接收到所述模型而未造成令人不满的延迟，那么语音识别服务可使用所述更大程度上自定义的额外模型，即使已经用另一模型开始或完成语音识别处理也是如此。

[0021] 在一些实施例中，内容服务器可对话语进行再处理（例如，多遍次 ASR 系统经配置

以对单个话语执行多次语音识别）。语音识别服务器或执行 ASR 的某一其它装置可具有至少一组可用的基础语音识别模型，或者可具有少量可用的额外选择（例如，性别特性的语音识别模型）。在用可用的模型（例如，基础模型）执行第一遍语音识别处理之后，可进行第二遍（如果及时检索到额外模型的话）。如果在第一遍之后没有返回额外或更特定的额外语音识别模型，那么结果可被返回到客户端装置。

[0022] 对于很多更大的语音识别模型（例如，语言模型）而言，可能难以足够快地检索到额外模型，因而无法将其用于实时语音识别。高速缓存额外的语音识别模型允许更快地检索到它们。例如，任何用户特定或以其它方式自定义的额外语音识别模型可存储在数据存储区中，所述数据存储区的容量较大，但响应时间相对较慢。另一数据存储区可用作更快返回额外语音识别模型的高速缓存。高速缓存可基于最近最少使用（“LRU”）标准而使额外的语音识别模型失效。

[0023] 当用户对着客户端装置说话时，ASR 系统可主动从高速缓存请求用户特定语音识别模型。如果高速缓存未中，那么 ASR 系统可继续使用最佳可用模型（例如，存储在语音识别服务器上的基础模型，或高速缓存中可用的不同额外模型）。高速缓存未中将导致用户特定语音识别模型被添加到高速缓存。由于用户通常使用 ASR 系统在短时间内处理多个话语（例如，两个或两个以上话语的语音识别会话），因此，任何检索到的用户特定模型都可用于除第一交互以外的所有交互。

[0024] 此外，分布式 ASR 系统可记录关于用户与分布式 ASR 系统的交互的数据。此类数据可用来检测模式和 / 或预测用户可能会使用分布式 ASR 系统的时间。用户特定或其它额外的语音识别模型可以预先高速缓存，从而它们在预测的时间可用。例如，用户可能会在每个工作日上午 8:00 左右开车上班的时候使用分布式 ASR 系统。在检测此类模式之后，分布式 ASR 系统可主动将用于用户的额外模型高速缓存在语音识别服务器上或高速缓存在网络高速缓存服务器中（例如，在 7:30 或 7:55 的时候）。当用户在上午 8:00 左右开始语音识别会话时，额外的模型将立即可用，并且可用来比基础模型更准确地处理第一话语，而没有等待时间或以其它方式与额外的语音识别模型相关联的检索延迟。

[0025] 在一些情况下，可检索自定义统计数据和模型的部分并使用其针对特定用户自定义基础声学或语言模型，而不是检索整个声学模型或语言模型。例如，分布式 ASR 系统可使用约束最大似然线性回归（“CMLLR”）变换、声道长度归一化（“VTLN”）扭曲因子、倒谱均值和方差、用于内插多个模型的权重和向量等等。有利的是，就将要传输的数据量而言，模型的这些部分通常比其可一起使用的声学或语言模型要小。因此，与检索整个额外的声学或语言模型相比，对模型的自定义或专用部分进行检索可大大减少检索时间，并且较少地影响用户所感知性能，而同时与使用基础模型可达到的结果相比，可提供更准确的结果。

[0026] 此外，语音识别模型及模型的部分可使用语音识别处理结果进行更新或进一步自定义。为了更新模型，可能会需要大型数据集。类似于语音识别模型的异步检索，可实施大型数据集的异步检索，以便在不影响用户所感知性能的情况下获取数据集。一旦已经检索到数据集，那么它们便可被用于更新额外的语音识别模型及模型的部分。另外，最新更新的模型可立即用来处理或再处理话语，这取决于系统要求和用户性能期望。

[0027] 在一些实施例中，替代于 ASR 模型或作为其补充，本文中描述的技术可用来检索额外的专用或自定义自然语言理解（“NLU”）模型。直观地说，与使用基础 NLU 模型处理文

本异步或并行，或者在 NLU 处理之前进行的 ASR 处理期间，可请求额外的 NLU 模型（例如，意图模型、所命名实体模型以及地名录）。由于额外的 NLU 模型被检索或以其它方式变得可用，因此，它们可用来重新计算 NLU 结果或可在随后的 NLU 处理期间使用。

[0028] 分布式 ASR 系统环境

[0029] 在详细描述用于管理分布式 ASR 系统中的额外语音识别模型的使用的过程的实施例之前，将描述可实施这些过程的示例性环境。图 1 示出了网络环境，所述网络环境包含客户端装置 102、ASR 服务器 104，以及 ASR 模型存储服务器 106。

[0030] 客户端装置 102 可对应于各种电子装置。在一些实施例中，客户端装置 102 可以是移动装置，该移动装置包含一个或多个处理器以及可含有处理器执行的软件应用的存储器。客户端装置 102 可含有麦克风或其它音频输入装置，用于接受语音输入，对所述语音输入将执行语音识别。直观地说，客户端装置 102 可以是移动电话、个人数字助理（“PDA”）、移动游戏装置、媒体播放器、电子书阅读器、平板计算机、膝上型计算机等。客户端装置 102 的软件可包含用于通过无线通信网络建立通信或直接与其它计算装置建立通信的部件。

[0031] ASR 服务器 104 可对从客户端装置 102 接收的用户话语执行自动化语音识别。ASR 服务器 104 可以是经配置以通过通信网络进行通信的任何计算系统。例如，ASR 服务器 104 可包含大量的服务器计算装置、台式计算装置、主计算机等。在一些实施例中，ASR 服务器 104 可包含物理上或逻辑上分组在一起的若干个装置，例如，经配置以对话语执行语音识别的应用服务器计算装置，以及经配置以存储记录和语音识别模型的数据库服务器计算装置。在一些实施例中，ASR 服务器 104 可包含组合在单个装置上的各种模块和部件、单个模块或部件的多个实例等。

[0032] 图 1 中示出的 ASR 模型存储服务器 106 可对应于一个或多个计算装置的逻辑关联，用于存储语音识别模型和通过网络为对于模型的请求提供服务。例如，ASR 模型存储服务器 106 可包含数据库服务器或存储部件，其对应于一个或多个服务器计算装置，用于获取和处理对于来自 ASR 服务器 104 的语音识别模型的请求。

[0033] ASR 服务器 104 可通过通信网络与客户端装置 102 和 / 或 ASR 模型存储服务器 106 通信。所述网络可以是可能由各个不同方操作的链接网络的公开可访问网络，例如因特网。在其它实施例中，所述网络可包含专用网络、个域网、局域网、广域网、电缆网络、卫星网络等或其某一组合，其中每个都可与因特网进行来回访问。例如，ASR 服务器 104 和 ASR 模型存储服务器 106 可位于单个数据中心内，并且可通过专用网络（例如，公司网络或校园网络）进行通信。客户端装置 102 可通过因特网与 ASR 服务器 104 通信。客户端装置 102 可通过有线或 WiFi 连接或者通过蜂窝电话网（例如，长期演进或 LTE 网）访问因特网。在一些实施例中，客户端装置 102 可与 ASR 模型存储服务器 106 直接通信。

[0034] 在一些实施例中，分布式 ASR 系统提供的特征和服务可实施为可通过通信网络消费的网络服务。在另外的实施例中，分布式 ASR 系统由实施在代管计算环境中的一个或多个虚拟机提供。代管计算环境可包含一个或多个快速配置和释放的计算资源，所述计算资源可包含计算装置、联网装置和 / 或存储装置。代管计算环境也可称为云计算环境。

[0035] 在操作过程中，客户端装置 102 可在 (A) 处接收来自用户的语音输入。客户端装置 102 可执行应用软件，所述应用软件可被用户激活，以接收声音输入。客户端装置 102 可通过集成麦克风、音频输入插孔或一些其它音频输入接口来接收声音输入。在一些实施例

中，当用户开始说话时，客户端装置 102 可自动接受声音输入，甚至用户无需激活声音输入或语音识别特征。

[0036] 在 (B) 处，客户端装置 102 将关于音频输入的音频信号或语音数据发送到 ASR 服务器 104。例如，客户端装置 102 可通过因特网直接与 ASR 服务器 104 或分布式 ASR 系统的某一其它部件（例如，管理部件）建立连接。直观地说，在具有多个 ASR 服务器 104 的分布式 ASR 系统中，管理部件可被实施以平衡多个 ASR 服务器 104 上的处理负载。在语音识别会话的持续期间，客户端装置 102（或其用户）可被分配到或以其它方式连接到特定的 ASR 服务器 104。在一些实施例中，语音识别会话可包含多个话语，所述多个话语被传输到 ASR 服务器 104，以用于在给定时间周期内或在紧密时间接近度内进行处理。

[0037] 在接收到声音输入或关于声音输入的数据之后，ASR 服务器 104 可在 (C) 处开始从 ASR 模型存储服务器 106 检索各种额外的语音识别模型。例如，ASR 服务器 104 可访问或接收关于说出话语的用户的数据，例如，用户的性别、地域口音等。所述数据可存储在用户简档中、与语音数据一起传输，或以某种其它方式获取。ASR 服务器 104 随后可确定一个或多个额外的语音识别模型或统计数据集，其可用来生成比基础模型或 ASR 服务器 104 可快速使用的其它模型更准确的结果。针对额外语音识别模型的请求可传输到 ASR 模型存储服务器 106。

[0038] 在 (D) 处，ASR 服务器 104 可开始使用 ASR 服务器 104 当前可用的最佳语音识别模型、统计数据以及其它数据来处理话语。在一些实施例中，ASR 服务器 104 可存储基础语音识别模型。此外，ASR 服务器 104 可存储经常使用的各种额外模型，例如，基于性别的模型。语音识别模型可消耗大量的存储空间，因此，在典型的实施方案中，ASR 服务器 104 可只存储少量的最基础模型或频繁使用的模型。

[0039] 当 ASR 服务器 104 使用其当前可用的模型来处理话语时，在 (C) 处请求的额外模型和其它数据可在 (E) 处被接收。由于 ASR 服务器 104 异步请求额外模型并且继续用其它模型开始处理，因此，额外模型可在 ASR 服务器 104 用其它模型完成处理话语期间或之后到达。

[0040] 一旦额外的模型被接收之后，它们可被用来在 (F) 处再处理初始结果或对其重新计分。在很多情况下，由于第一遍处理可能会减少重新计分的可能结果，因此，重新计分可比初始的第一遍语音识别处理执行得明显更快。因此，可用更适用的模型对初始结果进行重新计分，而不会将大量的延迟或等待时间添加到话语的整体语音识别处理。如果确定重新计分将造成性能令人不满或将不会显著提高结果的准确性，或者如果没有从 ASR 模型存储服务器 106 及时接收到模型，那么初始结果可在 (G) 处被传输到客户端装置 102。否则，如果结果被重新计分，那么重新计分的结果可在 (G) 处被传输到客户端装置 102。

[0041] 现转到图 2，将描述说明性 ASR 服务器 104。ASR 服务器 104 可包含 ASR 引擎 140、ASR 模型更新模块 142、管理模块 144、基础模型数据存储区 146，以及模型高速缓存 148。ASR 服务器 104 的每个模块、部件和数据存储区均可实施为单独的装置，或者各个个别模块、部件以及数据存储区可用各种组合的形式组合成单一装置。

[0042] ASR 引擎 140 可接收输入（例如，音频输入流或关于说出的话语的数据），并且使用各种语音识别模型和其它数据来确定最有可能的话语转录或其列表，如所属领域的技术人员将了解。ASR 模型更新模块 142 可使用来自 ASR 引擎 140 的结果和其它数据来更新可

用于生成更准确结果的额外模型及模型的部分。例如，ASR 引擎 140 可使用在多个语音识别会话的过程中发展的模型的一组用户特定或以其它方式自定义的部分。直观地说，ASR 引擎 140 使用的模型的部分可包含约束最大似然线性回归（“CMLLR”）变换、声道长度归一化（“VTLN”）扭曲因子、倒谱均值和方差、用于内插多个模型的权重和向量等等。有利的是，与完整的额外语音识别模型（例如，语言模型或声学模型）相比，模型的此些部分在存储、传送和使用期间消耗相对少量的空间、带宽、处理容量以及其它资源。此外，与单独使用基础语音识别模型相比，模型的此些部分还提高了语音识别过程的准确性。

[0043] 基于最近的结果更新模型和模型的部分可要求访问大型数据集（例如，计算声学模型所依据的基础数据集）。在 ASR 引擎 140 进行语音识别处理的过程中或之后，ASR 模型更新模块 142 或者 ASR 服务器 104 的某一其它模块或部件可异步检索大型数据集。当已接收到数据集时，其可用来更新额外的用户特定或以其它方式自定义的模型及模型的部分。同时，模型及模型的部分可在 ASR 处理期间继续被使用。

[0044] 管理模块 144 可监测 ASR 引擎 140 的进程以及额外语音识别模型的检索。如果管理模块 144 确定等待接收额外模型（或模型的部分）不会造成令人不满的性能延迟，那么管理模块 144 可致使 ASR 引擎 140 直到 ASR 引擎 140 有机会用额外模型对结果重新计分才将结果提供到客户端装置 102。然而，如果管理模块 144 确定等待接收额外模型会造成令人不满的性能延迟或不会显著提高结果的准确性，那么管理模块 144 可允许初始结果被提供到客户端装置 102 作为最终结果。

[0045] 基础模型数据存储区 146 可存储 ASR 引擎 140 在缺少更大程度上自定义、专门化或在其它方面更准确的额外模型时使用的基础声学模型和语言模型。此类基础模型可通过用户特定统计数据和模型的部分进行自定义，以提供更准确的结果。在一些实施例中，一个或多个最常用或广泛适用的额外模型（例如，性别特定模型）可存储在 ASR 服务器 104 的基础模型数据存储区 146 中，从而在需要的时候就无需从单独的 ASR 模型存储服务器 106 中进行检索。

[0046] 模型高速缓存 148 可用来存储被检索用于语音识别处理的额外模型和数据。例如，高速缓存可经配置以存储预先确定或动态确定量的数据。高速缓存可存储尽可能多的最近检索的模型，同时删除那些最近未被使用或请求的模型、使其失效或将其释放，以便为最新接收的模型腾出空间。各种高速缓存技术均可应用于模型高速缓存 148，包含使用存活时间（“TTL”）和最近最少使用（“LRU”）标准。

[0047] 管理模型检索的过程

[0048] 现在参考图 3，将描述用于管理语音识别模型的异步检索以及这些模型的使用的示例性过程 300。有利的是，ASR 服务器 104 可使用过程 300 来利用额外的语音识别模型及其它数据，如此将提高语音识别结果的准确性，而不会不利地影响所感知性能。

[0049] 过程 300 在框 302 处开始。在开始 ASR 会话之后，过程 300 可自动开始。过程 300 可体现为存储在分布式 ASR 系统的计算系统（例如，负载平衡管理器或个别 ASR 服务器 104）的计算机可读介质上（例如，一个或多个磁盘驱动器）的一组可执行程序指令。当过程 300 开始时，可执行程序指令可加载到存储器（例如，RAM）中并由计算系统的一个或多个处理器执行。

[0050] 在框 304 处，ASR 会话可被分配到特定的 ASR 服务器 104。由于额外语音识别模型

的检索，来自同一用户或同一客户端装置 102 针对 ASR 处理的随后请求可传输到同一 ASR 服务器 104，直到 ASR 会话结束的时间（例如，在经过一段时间之后或发生某一其它触发事件之后）为止。ASR 服务器 104 可访问或获取关于用户的数据，例如，用户的性别、年龄、地域口音或用户的身份。使用此人口统计数据或身份数据，ASR 服务器 104 可在框 306 处开始额外语音识别模型的检索。在一些实施例中，如上所述，与完整的额外语音识别模型相比，ASR 服务器 104 可针对当前用户检索模型的部分。在此类情况下，ASR 服务器 104 也可在框 320 处开始检索数据集，所述数据集可用来基于 ASR 处理的结果更新模型及模型的部分。在一些实施例中，用来更新模型部分的数据检索与 ASR 处理异步进行，从而使得在资源可用于进行此操作时以及在此检索和更新不妨碍 ASR 会话的处理时，检索和更新数据集。

[0051] 在决策框 308 处，ASR 服务器 104 可确定所请求的额外语音识别模型是否可立即使用。例如，所请求的模型可能在模型高速缓存数据存储区 148 中或在分布式 ASR 系统的单独模型高速缓存服务器中可用，如下文详细描述。在此类情况下，在框 314 处，高速缓存的额外模型可在初始 ASR 处理过程中被访问并使用，不论使用还是不使用 ASR 服务器 104 可用的任何基础模型（例如，基础模型数据存储区 146 中的模型）。如果没有额外的语音识别模型可用，或者如果 ASR 服务器 104 将仍然使用基础语音识别模型，那么在框 310 处，ASR 服务器 104 可在第一遍 ASR 处理过程中使用基础模型。在一些实施例中，所请求的额外语音识别模型可被高速缓存，但由于从高速缓存中检索模型的等待时间的缘故，ASR 服务器 104 将使用基础语音识别模型。

[0052] 在用基础语音识别模型进行第一遍 ASR 处理之后到达的决策框 312 处，ASR 服务器 104 可确定额外模型是否变得可用。若是，过程 300 可行进到框 314，其中 ASR 服务器 104 可用额外的语音识别模型执行第二遍 ASR（例如，对初始结果进行重新计分）。此外，任何最近接收的额外语音识别模型均可被高速缓存。

[0053] 否则，如果额外模型尚未被接收到，或者如果确定使用额外模型将造成令人不满的性能延迟或不能显著提高准确性，那么过程 300 可行进到框 316。

[0054] 在框 316 处，ASR 服务器 104 可将最终结果传输到客户端装置 102。在一些实施例中，ASR 服务器 104 可执行某一动作或致使另一装置执行动作，而不是将结果传输到客户端装置 102。例如，来自 ASR 过程的结果可以提供到自然语言理解（“NLU”）部件，所述部件经配置以根据用户话语来确定用户意图。基于用户意图（例如，找方向、订航班、开始声音拨号），ASR 服务器 104 可执行某一动作。

[0055] 在将结果发送到客户端装置 102（或导致执行某一其它动作）之后，在决策框 318 处，ASR 服务器 104 可在同一 ASR 会话期间等待额外话语以便处理。如果另一话语被接收到，那么过程 300 可返回到框 306。否则，如果在一段时间内未接收到另一话语，或者如果发生另一触发事件（例如，例如通过将客户端装置 102 断电，用户肯定地结束了 ASR 会话），那么过程 300 可在框 324 处结束。

[0056] 除了等待额外的话语之外，ASR 模型更新模块 142 或 ASR 服务器 104 的某一其它部件还可在框 322 处基于 ASR 处理的结果来更新数据集。更新过程可利用在框 320 处异步检索的数据集。经更新的数据集随后可被高速缓存、传输到 ASR 模型存储服务器 106、在第二遍 ASR 处理过程中使用等。在一些实施例中，只要用于更新数据集的 ASR 结果可用，额外的模型或模型的部分便可基于经更新的数据集进行更新或重新计算，例如，与框 316 并行

进行或在框 314 之后即刻进行。

[0057] 用于高速缓存模型的过程和结构

[0058] 现在转到图 4，将描述基于用户活动的预测来预先高速缓存额外语音识别模型的示例性过程 400。有利的是，过程 400 可用来分析先前的用户活动、预测用户可能利用分布式 ASR 系统的时间，以及预先高速缓存额外的模型，从而它们准备好在所预测时间立即或大体上立即使用。

[0059] 过程 400 在框 402 处开始。过程 400 可在 ASR 服务器 104 或分布式 ASR 系统的某一其它部件加电之后自动开始，或者其可以手动开始。过程 400 可体现为存储在与分布式 ASR 系统相关联的计算系统的计算机可读介质（例如，一个或多个磁盘驱动器）上的一组可执行程序指令。当过程 400 开始时，可执行程序指令可加载到存储器（例如，RAM）中并由计算系统的一个或多个处理器执行。

[0060] 在框 404 处，分布式 ASR 系统可处理语音识别会话，如上所述。在框 406 处，关于特定用户的 ASR 会话的使用数据可在处理 ASR 会话的时候记录下来。例如，代管语音识别会话的 ASR 服务器 104 的管理模块 144 可记录关于用户或客户端装置 102 的数据，包含 ASR 请求的日期和时间、结果的内容、请求的主题或上下文等等。

[0061] 在框 408 处，管理模块 144 或分布式 ASR 系统的某一其它模块或部件可检测所记录数据中的模式，或确定关于用户可能会访问分布式 ASR 系统的时间的预测。例如，特定用户可能会在工作日上午 8:00 或左右，有规律地向分布式 ASR 系统传输语音数据以用于处理。分布式 ASR 系统的部件可检测此类模式，并且作为响应，预测用户将在下一个工作日上午 8:00 再次传输语音数据。此外，用户可在这些上午 8:00 的会话期间依照惯例传输关于全球定位系统（“GPS”）方向或音乐回放的声音命令。通过包含此类细节，可使预测更加具体。可基于详细预测高速缓存以此类活动为目标的额外语音识别模型。

[0062] 在框 410 处，预期用户会开始 ASR 会话，分布式 ASR 系统可以在下一个工作日上午 8:00 之前不久为用户预先高速缓存额外的模型。例如，在用户开始 ASR 会话之前，用户可在上午 7:55 或上午 7:59 被主动分配到特定的 ASR 服务器 104。用于用户的额外模型可被预先高速缓存在所分配的 ASR 服务器 104 处，从而在用户开始会话时可以立即使用。例如，模型可以存储在 ASR 服务器 104 的模型高速缓存 148 中。被选择用于预先高速缓存的模型的选择依据可以是 :用户的人口统计数据或身份、所预测的会话的主题、它们的某一组合等等。在一些实施例中，额外的模型可以高速缓存在 ASR 模型存储服务器 106 与 ASR 服务器 104 之间的中间高速缓存处，如下文详细描述。在此类情况下，由于多个服务器可从中间高速缓存检索高速缓存的模型，因此，用户可以不被主动分配到特定的 ASR 服务器 104。

[0063] 在一些实施例中，计算额外模型将被高速缓存的时间可以基于用户先前访问时间的分布，而非特定的平均值或可能访问时间的预计。所述计算可以使得所选择的时间将导致额外的模型在某一时间被高速缓存，所述某一时间在用户先前或预计访问时间的阈值量或百分数之前。返回到上述实例，用户可通常在 8:00 左右开始 ASR 会话，但实际的时间分布可以从上午 7:30 延伸到上午 8:30。管理模块 144 可确定在上午 7:30 高速缓存额外的模型，且在该时间将用户分配到特定的 ASR 服务器 104 将导致额外模型可用于 90％或 99％的用户的“上午 8:00”ASR 会话。

[0064] 在框 412 处，用户可通过分布式 ASR 系统开始 ASR 会话。负载平衡部件或分布式

ASR 系统的某一其它部件可确定 :用户已经与用于会话的特定 ASR 服务器 104 相关联，并且如果高速缓存尚未失效，或者如果在阈值时间段已经过去之后用户没有意外开始 ASR 会话，那么话语数据可被发送到主动分配的 ASR 服务器 104。例如，如果用户在上午 7:30 到上午 8:30 之间开始会话，那么用户可连接到主动分配的 ASR 服务器 104，并实现预先高速缓存带来的好处。然而，如果用户直到上午 9:00 才开始会话，或者如果高速缓存的模型已经被释放以为最近请求或使用的模型腾出空间，那么用户的 ASR 会话可处理作为任何其他用户的 ASR 会话，例如，如上文参考图 3 所描述。

[0065] 在一些实施例中，基于最近用户交互或环境因素，语音识别模型可以肯定地加载或预先高速缓存。例如，客户端装置 102 可以监测来自麦克风的输入，并且经配置以识别用户说出的某个单词或短语，以在不与装置物理交互（例如，不用按下按钮或与触摸屏交互）的情况下开始 ASR 会话。在一些情况下，当满足某些条件时（例如，初步分析表明，这是与环境噪音不同的话语），来自麦克风的音频输入可以传输到分布式 ASR 系统，以确定用户是否说出了指示开始 ASR 会话的单词或短语。在一些情况下，客户端装置 102 可监测一定空间中是否出现用户，因为用户进入空间后可能会很快对着客户端装置 102 说话。当客户端装置 102 检测到用户出现（例如，使用传感器、对视频信号使用图像处理，或对音频信号使用信号处理）时，消息可被发送到分布式 ASR 系统，以表明用户可能会很快通过客户端装置 102 开始语音识别。在这些和其它情况下，在完整的话语被传输到分布式 ASR 系统以用于处理之前，可针对用户加载额外的语音识别模型。

[0066] 图 5A 和图 5B 示出用于分布式 ASR 系统 110 中的多层 ASR 模型存储和高速缓存的说明性结构。分布式 ASR 系统 110 可包含多个 ASR 服务器 104a、104b、长期 ASR 模型存储服务器 106，以及高速 ASR 模型高速缓存 108。以物理接近度测量或就交换通信所需的时间量或网络跃点数目而言，比起接近 ASR 模型存储服务器 108，ASR 服务器 104a、104b 可以更接近高速 ASR 模型高速缓存 108。此外，与 ASR 模型存储服务器 106 相比，ASR 模型高速缓存 108 可利用不同的硬件来提供更快的性能，但容量更少。在一些实施例中，分布式 ASR 系统 110 可包含多个 ASR 模型高速缓存 108，例如，用于每 n 个 ASR 服务器 104 的一个 ASR 模型高速缓存 108，其中 n 可以是任何数字。

[0067] 在 (A) 处，客户端装置 102a 可将语音数据发送到分布式 ASR 系统 110 以用于处理。在 (B) 处，针对额外语音识别模型的请求可以从 ASR 服务器 104a 发出到 ASR 模型高速缓存 108，而非发出到 ASR 模型存储服务器 106。如果 ASR 模型高速缓存 108 具有可用的所请求模型，那么高速缓存的模型可以返回到 ASR 服务器 104a，比从长期 ASR 模型存储服务器 106 检索语音识别模型明显更快。如果 ASR 模型高速缓存 108 不具有所请求的模型，那么 ASR 模型高速缓存 108 可在 (C) 处从 ASR 模型存储服务器 106 检索所请求的模型，在 (D) 处高速缓存语音识别模型的副本，并且在 (E) 处将副本转发到发出请求的 ASR 服务器 104a。ASR 模型高速缓存 108 可应用各种高速缓存技术，包含使用存活时间（“TTL”）和最近最少使用（“LRU”）标准。ASR 服务器 104a 可在 (F) 处将结果传输到客户端装置 102a 或基于 ASR 结果来执行某一动作。

[0068] 有利的是，针对最新高速缓存的语音识别模型的随后请求可从 ASR 模型高速缓存 108 得到服务，而不是从 ASR 模型存储服务器 108 得到服务。例如，客户端装置 102a 可将语音数据提交到同一 ASR 服务器 104a 或不同的 ASR 服务器 104b，并且在任一种情况下，额外

的语音识别模型均可从 ASR 模型高速缓存 108 检索，而无需从 ASR 模型存储服务器 106 检索。作为另一实例，语音数据可从不同的客户端装置 102b 接收，并由同一 ASR 服务器 104a 或不同的 ASR 服务器 104b 处理。如图 5B 所示，第二客户端装置 102b 可在 (G) 处将语音数据传输到分布式 ASR 系统 110。在 (H) 处，第二 ASR 服务器 104b 可处理语音数据，请求来自 ASR 模型高速缓存 108 的相同额外语音识别模型。由于先前已高速缓存了模型，因此，所请求的模型可在 (I) 处返回到 ASR 服务器 104B，而不用从 ASR 模型存储服务器 106 检索。ASR 服务器 104b 可在 (J) 处将结果传输到客户端装置 102b 或基于 ASR 结果来执行某一动作。

[0069] 术语

[0070] 根据实施例，本文所描述的过程或算法中的任一个的某些动作、事件或功能可用不同的序列来执行，可被添加、合并或完全忽略（例如，并非所有所描述的操作或事件都是实践算法所必需的）。此外，在某些实施例中，操作或事件可同时执行（例如，通过多线程处理、中断处理，或者多个处理器或处理器芯，或者在其它并行结构上），而非循序执行。

[0071] 结合本文中公开的实施例描述的各种说明性逻辑块、模块、例程和算法步骤可实施为电子硬件、计算机软件，或是这两个的组合。为了清楚地说明硬件和软件的这种互换性，上文就其功能性描述了各种说明性部件、块、模块和步骤。此类功能性是实施为硬件还是软件取决于特定应用和强加于整个系统的设计约束。所描述的功能性可针对每个特定应用用不同的方式实施，但此类实施决策不应被解释为导致脱离本发明的范围。

[0072] 结合本文公开的实施例描述的方法、过程、例程或算法的步骤可直接体现在硬件、处理器执行的软件模块或这两个的组合中。软件模块可驻留于 RAM 存储器、闪存、ROM 存储器、EPROM 存储器、EEPROM 存储器、寄存器、硬磁盘、可移动磁盘、CD-ROM 或任何其它形式的非临时计算机可读存储介质中。示例性存储介质可耦合到处理器，从而使得处理器可从存储介质中读取信息并将信息写入到存储介质。在替代方案中，存储介质可以与处理器成一体式。处理器和存储介质可驻留于 ASIC 中。ASIC 可驻留于用户终端中。在替代方案中，处理器和存储介质可以作为离散部件驻留在用户终端中。

[0073] 除非另外特别说明，或者根据所使用的上下文可用其它方式理解，否则本文中所用的条件性语言，例如，“可”、“可以”、“可能”、“也许”、“例如”等，大体意图传达以下内容：某些实施例包含（但其它实施例不包含）某些特征、元件和 / 或步骤。因此，这些条件性语言大体并不意图暗示：特征、元件和 / 或步骤是一个或多个实施例无论如何都需要的，或者在有或没有作者输入或提示的情况下，一个或多个实施例必然包含用于决策的逻辑，不论这些特征、元件和 / 或步骤均包含其中还是将在任何特定的实施例中执行。术语“包括”、“包含”、“具有”等是同义词，且在包含性意义上以开放的形式使用，并且并不排除额外的元件、特征、动作、操作等等。此外，术语“或”在包含性意义（而并非排除性意义）上使用，因此在用于（例如）连接一系列元件时，术语“或”表示所述系列中的元件的一个、一些或所有元件。

[0074] 除非另外特别说明，否则连接性语言（例如，短语“X、Y 和 Z 中的至少一个”）将在所使用的上下文中被理解为大体表示一个项目、术语等可以是 X、Y 或 Z 中任一者或者其组合。因此，这些连接性语言通常并不意图暗示某些实施例要求分别存在 X 中的至少一个、Y 中的至少一个以及 Z 中的至少一个。

[0075] 虽然上文的具体实施方式已展示、描述并指出应用于各种实施例的新颖特征，但

是应理解，在不脱离本发明的精神的情况下，可作出所说明的装置或算法的形式和细节上的各种省略、取代和改变。应认识到，本文描述的本发明的某些实施例可采用不提供本文所陈述的所有特征和益处的方式体现，因为一些特征可独立于其它特征单独使用或实践。本文公开的特定发明的范围由所附权利要求书指示，而非由上述实施方式指示。权利要求书的范围涵盖其等效意义和范围内的所有变化。

[0076] 条款

[0077] 1. 一种系统，其包括：

[0078] 存储可执行指令的计算机可读存储器；以及

[0079] 与所述计算机可读存储器通信的一个或多个处理器，其中所述一个或多个处理器经所述可执行指令编程以：

[0080] 从客户端装置接收包括用户话语的音频数据；

[0081] 确定额外语音识别模型不可用；

[0082] 使用基础语音识别模型来对所述音频数据执行第一语音识别处理，以产生第一语音识别结果；

[0083] 从网络可访问的数据存储区请求所述额外语音识别模型，其中所述请求是在完成所述第一语音识别处理之前开始的；

[0084] 从所述网络可访问的数据存储区接收所述额外语音识别模型；

[0085] 使用所述额外语音识别模型以及使用所述音频数据或所述语音识别结果中的至少一个来执行第二语音识别处理；以及

[0086] 至少部分基于所述第二语音识别处理，将响应传输到所述客户端装置。

[0087] 2. 根据条款 1 所述的系统，其中所述基础语音识别模型包括通用声学模型、性别特定声学模型或通用语言模型中的至少一个，并且其中至少部分基于与所述用户话语相关联的用户的特性来选择所述额外语音识别模型。

[0088] 3. 根据条款 1 所述的系统，其中所述一个或多个处理器还经所述可执行指令编程以：

[0089] 从所述客户端装置接收包括第二用户话语的第二音频数据；

[0090] 确定所述额外语音识别模型可用；以及

[0091] 使用所述额外语音识别模型对所述第二音频数据执行语音识别处理。

[0092] 4. 根据条款 1 所述的系统，其中所述一个或多个处理器还经所述可执行指令编程，以使用多线程处理以与所述第一语音识别处理的执行并行检索所述额外语音识别模型。

[0093] 5. 根据条款 1 所述的系统，其中所述一个或多个处理器还经所述可执行指令编程以高速缓存所述额外语音识别模型。

[0094] 6. 一种计算机实施的方法，其包括：

[0095] 在以特定计算机可执行指令配置的一个或多个计算装置的控制下，

[0096] 对关于用户话语的音频数据执行第一语音处理，以产生语音处理结果；

[0097] 从网络可访问的数据存储区请求语音处理数据，其中所述请求是在完成所述第一语音处理之前开始的；

[0098] 从所述网络可访问的数据存储区接收所述语音处理数据；以及

[0099] 使用所述语音处理数据以及所述音频数据或所述语音处理结果中的至少一个来执行第二语音处理。

[0100] 7. 根据条款 6 所述的计算机实施的方法，其还包括：

[0101] 至少部分基于所述用户的特性来选择待请求的语音处理数据。

[0102] 8. 根据条款 7 所述的计算机实施的方法，其中所述用户的所述特性包括所述用户的性别、年龄、地域口音或身份。

[0103] 9. 根据条款 6 所述的计算机实施的方法，其中所述语音处理数据包括以下至少一个：声学模型、语言模型、语言模型统计数据、约束最大似然线性回归（“CMLLR”）变换、声道长度归一化（“VTLN”）扭曲因子、倒谱均值和方差数据、意图模型、所命名实体模型或地名录。

[0104] 10. 根据条款 9 所述的计算机实施的方法，其还包括：

[0105] 请求用于更新所述语音处理数据的统计数据，其中针对统计数据的所述请求是在完成所述第一语音处理之前开始的。

[0106] 11. 根据条款 10 所述的计算机实施的方法，其还包括至少部分基于所述统计数据和所述第二语音处理的结果来更新所述语音处理数据。

[0107] 12. 根据条款 6 所述的计算机实施的方法，其还包括高速缓存所述语音处理数据。

[0108] 13. 根据条款 12 所述的计算机实施的方法，其还包括：

[0109] 接收关于所述用户的第二话语的第二音频数据；

[0110] 从高速缓存检索所述语音处理数据；以及

[0111] 使用所述语音处理数据对所述第二音频数据执行语音处理。

[0112] 14. 根据条款 12 所述的计算机实施的方法，其中高速缓存所述语音识别数据包括将所述语音识别数据存储在与所述网络可访问的数据存储区分开的高速缓存服务器处。

[0113] 15. 根据条款 12 所述的计算机实施的方法，其中检索所述语音识别数据包括检索所述语音识别数据的经高速缓存副本。

[0114] 16. 根据条款 12 所述的计算机实施的方法，其中高速缓存所述语音识别数据包括：

[0115] 确定所述用户可能会开始语音识别会话的时间；以及

[0116] 大体在所述所确定的时间高速缓存所述语音识别数据。

[0117] 17. 根据条款 6 所述的计算机实施的方法，其还包括：

[0118] 从所述用户操作的客户端装置接收所述音频数据；以及

[0119] 将响应传输到所述客户端装置，所述响应至少部分基于所述第二语音识别处理。

[0120] 18. 根据条款 6 所述的计算机实施的方法，其还包括：

[0121] 至少部分基于所述第二语音识别处理来执行动作。

[0122] 19. 一种包括可执行代码的非临时计算机可读介质，所述可执行代码在被处理器执行时致使计算装置执行过程，所述过程包括：

[0123] 对关于用户话语的音频数据执行第一语音识别处理，以产生语音识别结果；

[0124] 从网络可访问的数据存储区请求语音识别数据，其中所述请求是在完成所述第一语音识别处理之前开始的；

[0125] 从所述网络可访问的数据存储区接收所述语音识别数据；以及

[0126] 使用所述语音识别数据以及所述音频数据或所述语音识别结果中的至少一个来执行第二语音识别处理。

[0127] 20. 根据条款 19 所述的非临时计算机可读介质，其中所述过程还包括：

[0128] 至少部分基于接收到所述音频数据的日期或时间中的一个来选择待请求的语音识别数据。

[0129] 21. 根据条款 19 所述的非临时计算机可读介质，其中所述过程还包括：

[0130] 至少部分基于与所述用户相关联的特性来选择待请求的语音识别数据。

[0131] 22. 根据条款 21 所述的非临时计算机可读介质，其中与所述用户相关联的所述特性包括以下一个：所述用户的性别、年龄、地域口音、身份或与所述用户相关联的群组的身份。

[0132] 23. 根据条款 19 所述的非临时计算机可读介质，其中所述语音识别数据包括：声学模型、语言模型、语言模型统计数据、约束最大似然线性回归（“CMLLR”）变换、声道长度归一化（“VTLN”）扭曲因子、倒谱均值和方差数据、意图模型、所命名实体模型或地名录。

[0133] 24. 根据条款 23 所述的非临时计算机可读介质，其中所述过程还包括：

[0134] 请求用于更新所述语音识别数据的统计数据，其中针对统计数据的所述请求是在完成所述第一语音识别处理之前开始的。

[0135] 25. 根据条款 24 所述的非临时计算机可读介质，其中所述过程还包括：

[0136] 至少部分基于所述统计数据和所述第二语音识别处理的结果来更新所述语音识别数据。

[0137] 26. 根据条款 19 所述的非临时计算机可读介质，其中所述过程还包括：

[0138] 高速缓存所述语音识别数据。

[0139] 27. 根据条款 19 所述的非临时计算机可读介质，其中检索所述语音识别数据包括检索所述语音识别数据的经高速缓存副本。

[0140] 28. 根据条款 19 所述的非临时计算机可读介质，其中所述过程还包括：

[0141] 从所述用户操作的客户端装置接收所述音频数据；以及

[0142] 将响应传输到所述客户端装置，所述响应至少部分基于所述第二语音识别处理。

[0143] 29. 根据条款 19 所述的非临时计算机可读介质，其中所述过程还包括：

[0144] 至少部分基于所述第二语音识别处理来执行动作。

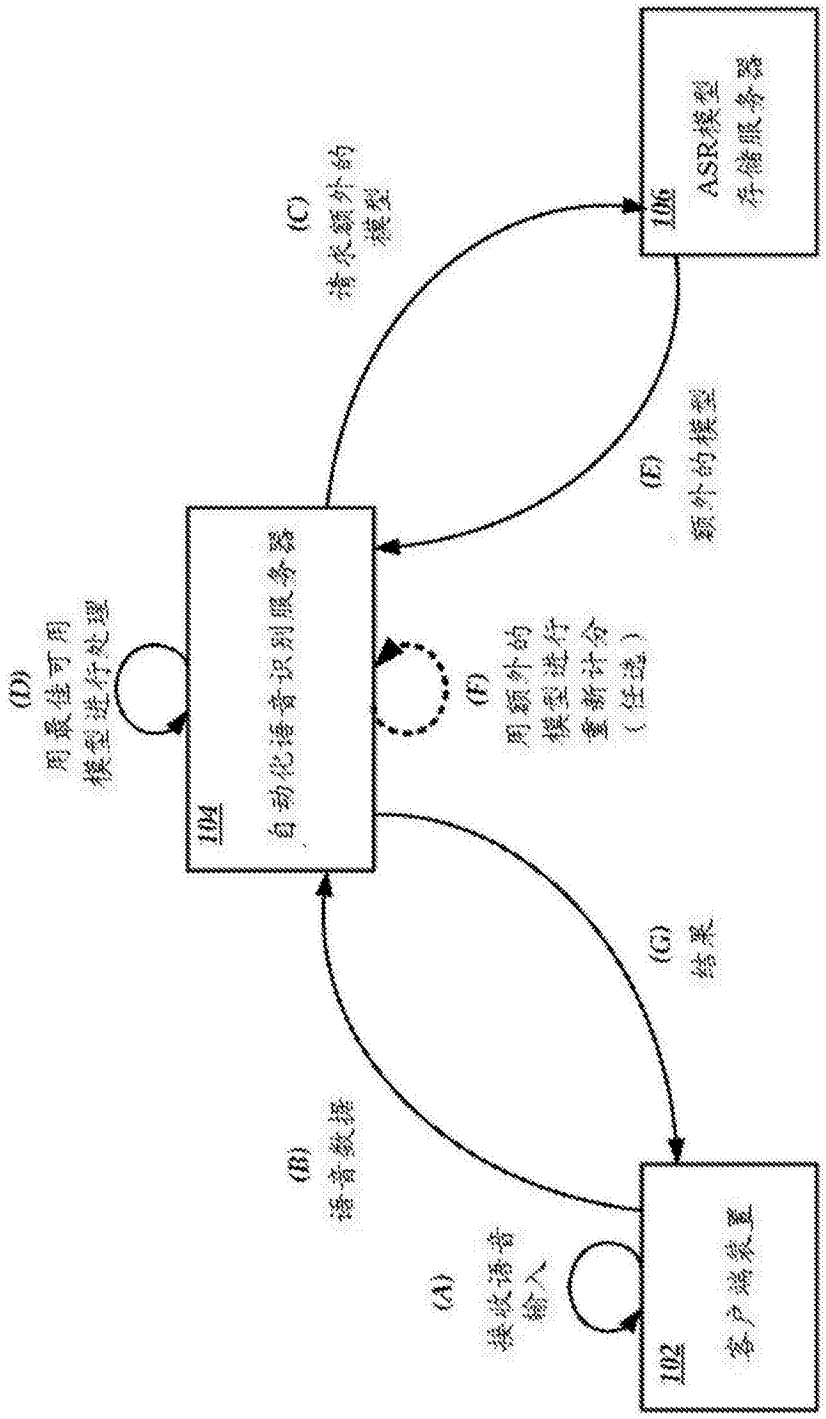

102
客户端装置
104
自动化语音识别服务器
106
ASR模型
存储服务器
(A)
接收语音
输入
(B)
语音数据
(C)
请求额外的
模型
(D)
用最佳可用
模型进行处理
(E)
额外的模型
(F)
用额外的
模型进行
重新计分
(任选)
(G)
结果

图 1

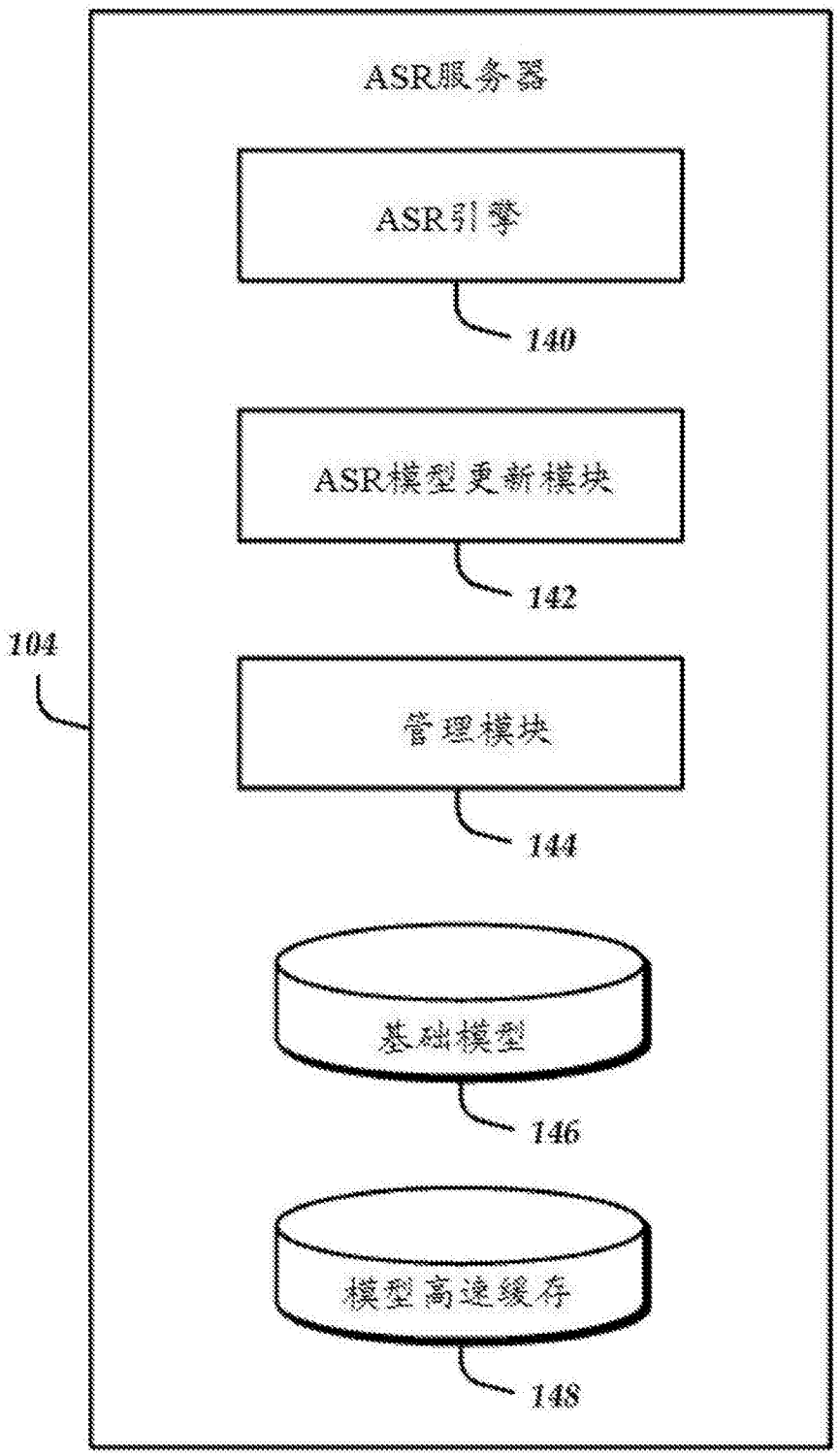

ASR服务器
ASR引擎
140
ASR模型更新模块
142
104
管理模块
144
基础模型
146
模型高速缓存
148

图 2

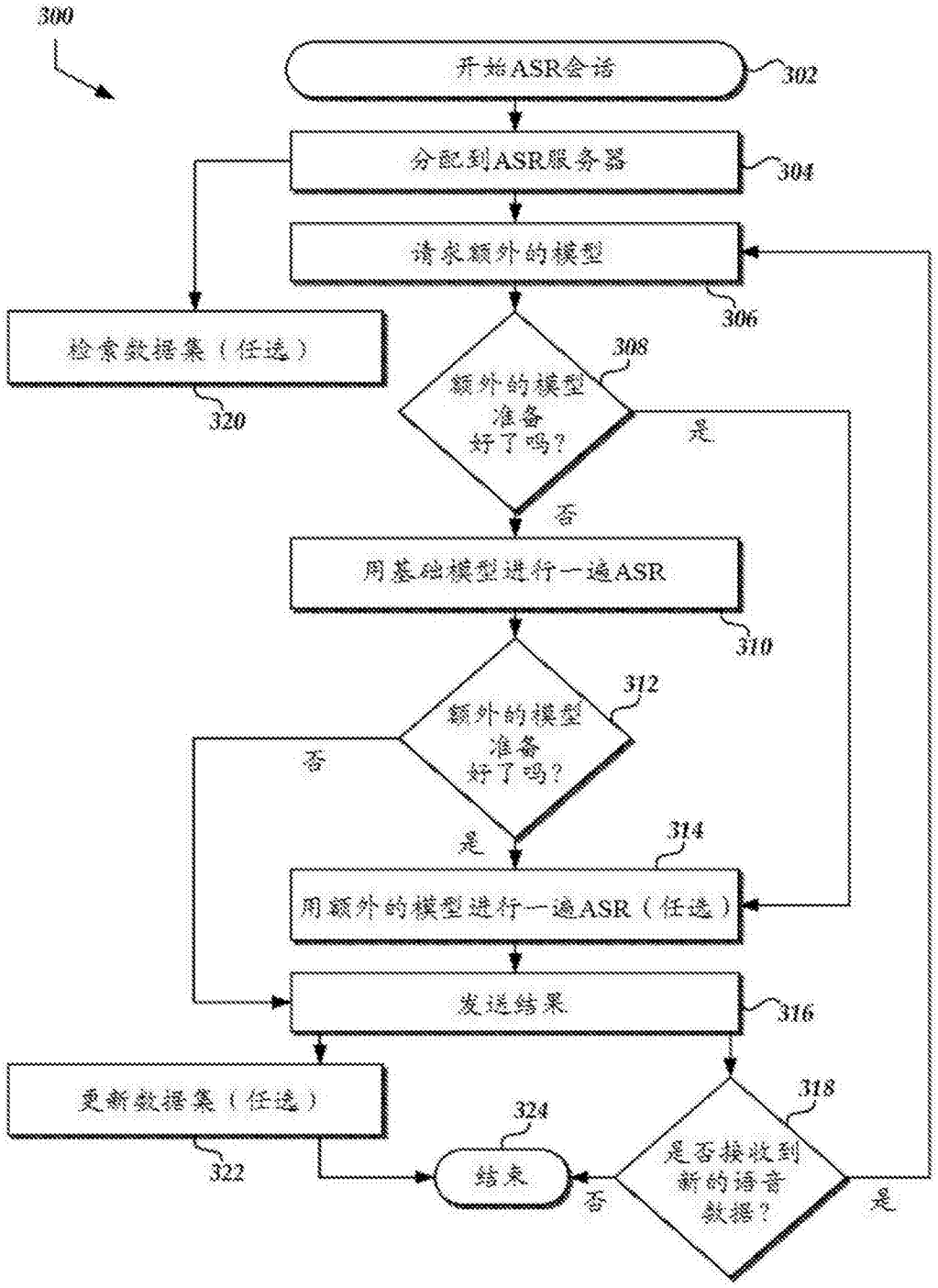

300
开始ASR会话
302
分配到ASR服务器
304
请求额外的模型
306
检索数据集（任选）
320
额外的模型准备好了吗？
308
是
否
用基础模型进行一遍ASR
310
额外的模型准备好了吗？
312
否
是
314
用额外的模型进行一遍ASR（任选）
发送结果
316
更新数据集（任选）
322
结束
324
是否接收到新的语音数据？
318
否
是

图 3

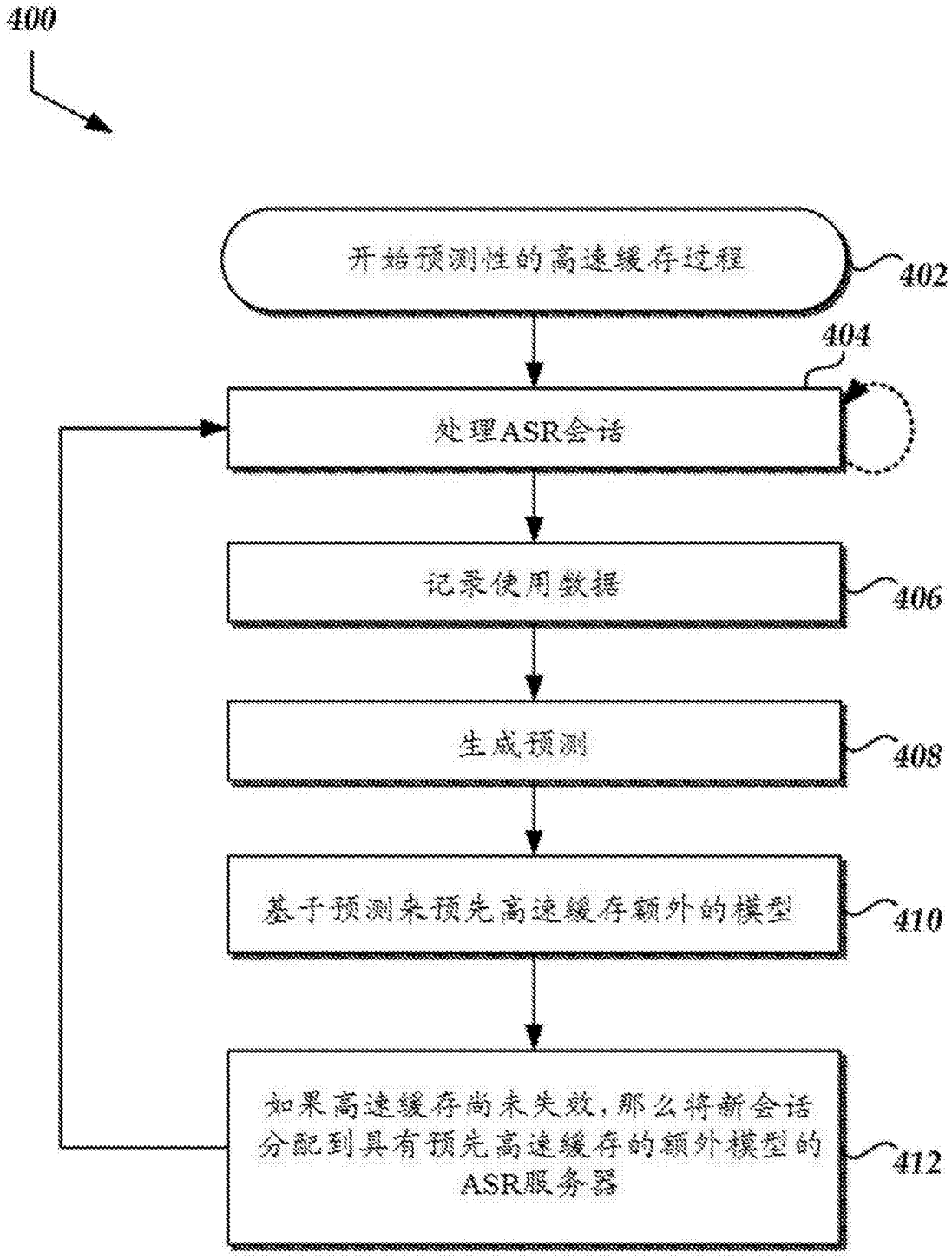

400
开始预测性的高速缓存过程
402
404
处理ASR会话
记录使用数据
406
生成预测
408
基于预测来预先高速缓存额外的模型
410
如果高速缓存尚未失效，那么将新会话分配到具有预先高速缓存的额外模型的ASR服务器
412

图 4

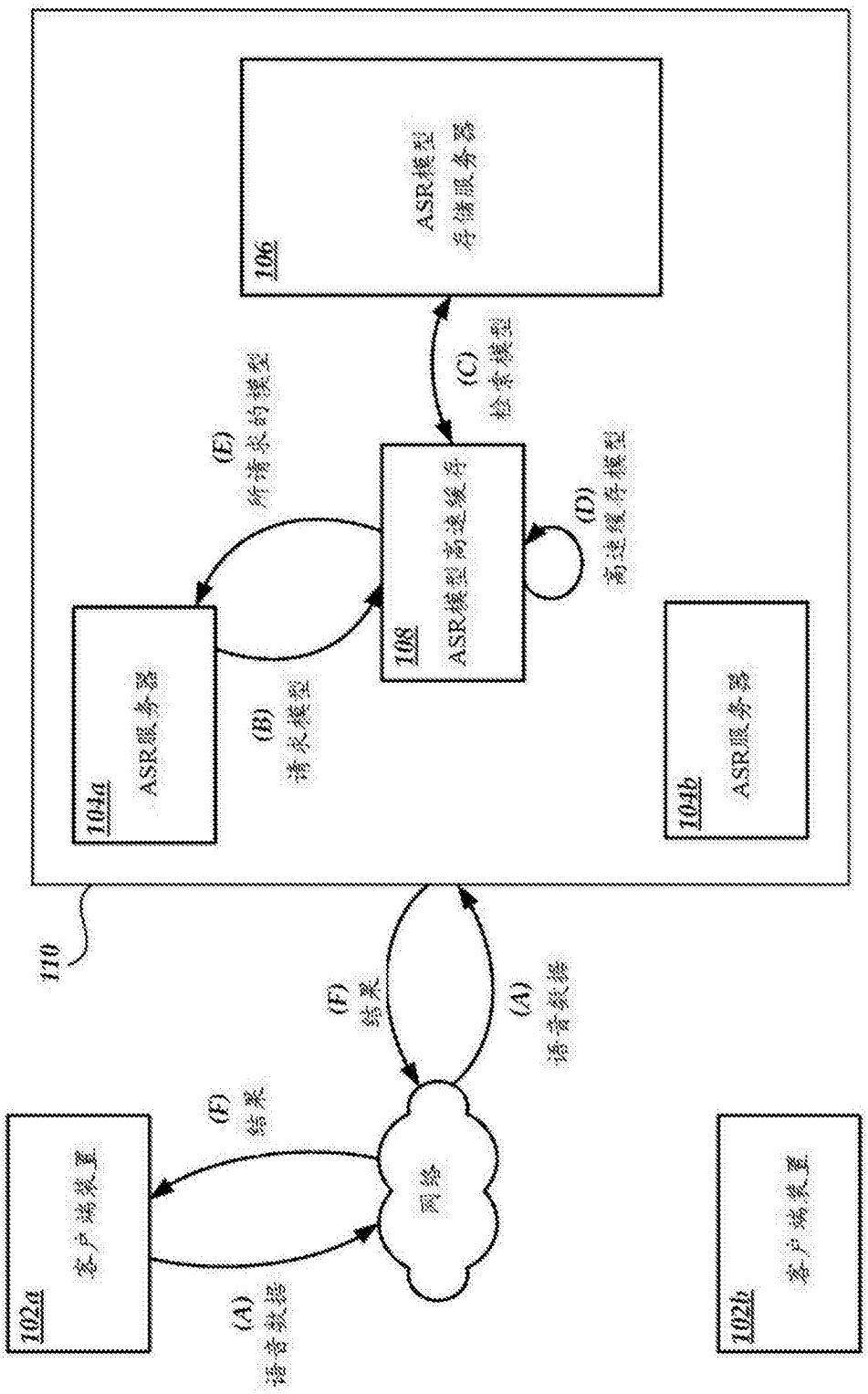

106
ASR模型
存储服务器
(C)
检索模型
(E)
所请求的模型
108
ASR模型高速缓存
(D)
高速缓存模型
104a
ASR服务器
(B)
请求模型
104b
ASR服务器
110
(F)
结果
(A)
语音数据
(F)
结果
网络
102a
客户端装置
(A)
语音数据
102b
客户端装置

图 5A

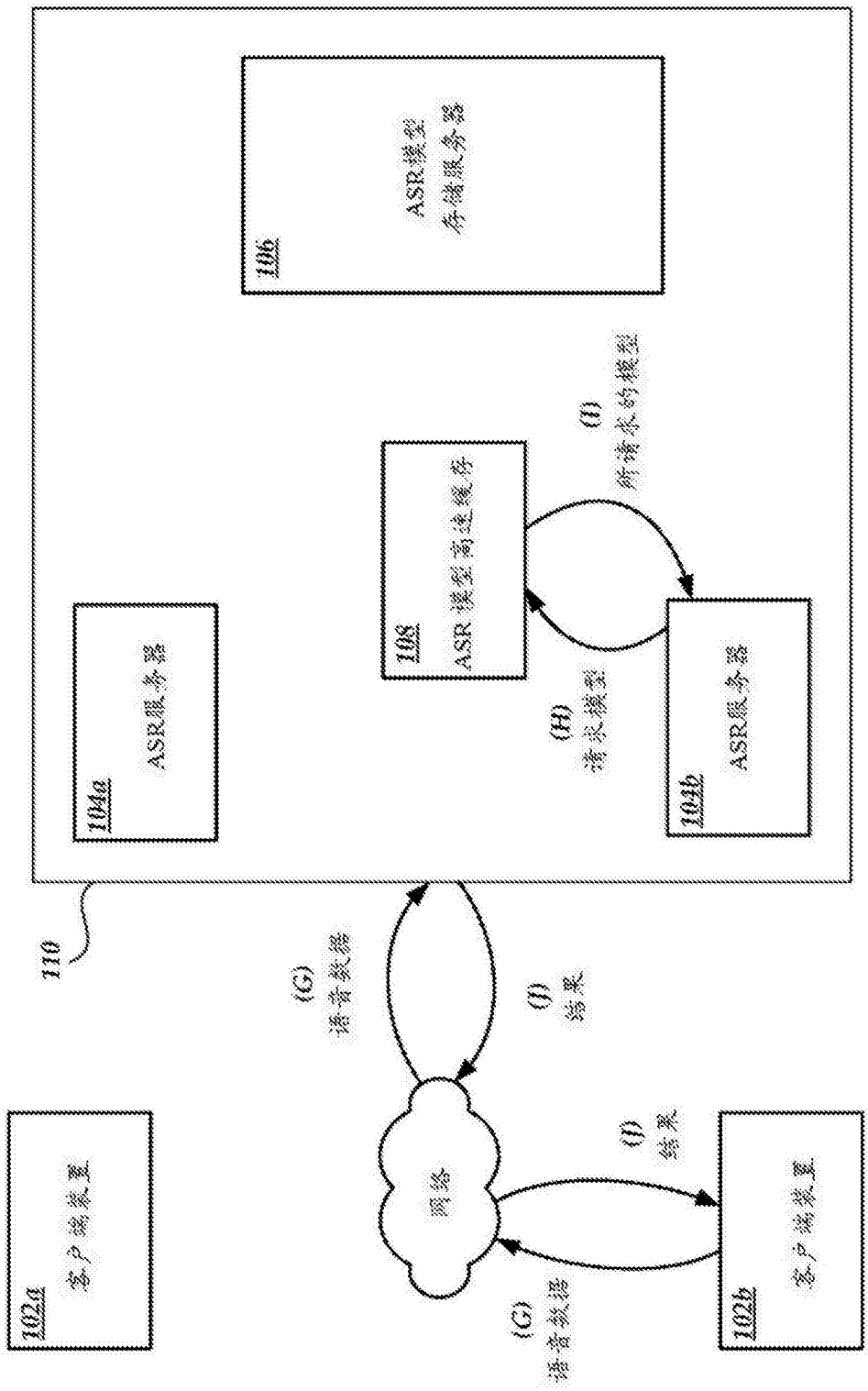

106
ASR模型
存储服务器
108
ASR 模型高速缓存
(I)
所请求的模型
(H)
请求模型
104a
ASR服务器
104b
ASR服务器
110
(G)
语音数据
(J)
结果
102a
客户端装置
网络
(J)
结果
(G)
语音数据
102b
客户端装置

图 5B