(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 112673421 A
(43)申请公布日 2021.04.16

(21)申请号 201980026087.0
(22)申请日 2019.11.27
(30)优先权数据
62/772,514 2018.11.28 US
62/772,922 2018.11.29 US
(85)PCT国际申请进入国家阶段日
2020.10.15
(86)PCT国际申请的申请数据
PCT/US2019/063643 2019.11.27
(87)PCT国际申请的公布数据
WO2020/113031 EN 2020.06.04
(71)申请人 谷歌有限责任公司
地址 美国加利福尼亚州
(72)发明人 万里 于洋 普拉尚特·斯里达尔 伊格纳西奥·洛佩斯·莫雷诺 王泉
(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 周亚荣 邓聪惠
(51)Int.Cl.
*G10L 15/00*(2013.01)
*G10L 15/16*(2006.01)
*G10L 15/183*(2013.01)
*G10L 15/14*(2006.01)
*G10L 25/24*(2013.01)
*G06N 3/00*(2006.01)

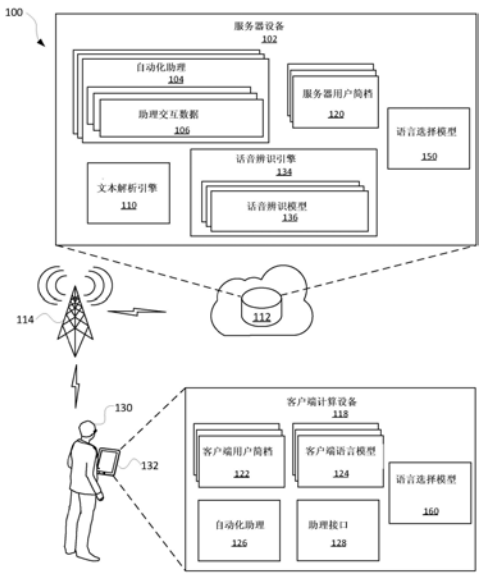
权利要求书3页 说明书18页 附图6页

(54)发明名称

# 训练和/或使用语言选择模型以自动确定用于口头话语的话音辨识的语言

(57)摘要

用于训练和/或使用语音选择模型以在确定音频数据中捕获的口头话语的特定语言时使用。可以使用经训练的语言选择模型处理音频数据的特征以生成N种不同语言中的每一种语言的预测概率,并且基于所生成的概率选择特定语言。可以响应于选择了口头话语的特定语言而采用针对该特定语言的话音辨识结果。许多实施方式涉及利用元组损失代替传统的交叉熵损失来训练语言选择模型。利用元组损失训练语言选择模型可以导致更加有效的训练和/或可以导致更加准确和/或鲁棒的模型——由此缓解了针对口头话语的错误语言选择。




CN 112673421 A

1.一种由一个或多个处理器实现的方法,所述方法包括:

生成多个训练示例,其中,生成所述训练示例中的每一个训练示例基于捕获相对应人类话语的相对应音频数据以及指示所述相对应人类话语的相对应口头语言的相对应标记,所述相对应口头语言是要辨识的N种不同语言中的一种语言,其中,N是大于10的整数,并且其中,训练示例中的每一个训练示例包括:

相对应的训练示例输入,所述相对应的训练示例输入包括:所述相对应音频数据的相对应特征;和

相对应的训练示例输出,所述相对应的训练示例输出包括:针对要辨识的所述N种不同语言的中的每一种语言的相对应标记概率量度,其中,所述相对应标记概率量度基于相对应标记包括对应于所述相对应口头语言的相对应正概率量度标记,以及针对所述相对应标记概率量度的所有其它相对应标记概率量度的相对应负概率量度标记;以及

基于所述训练示例训练语言选择模型,训练所述语言选择模型包括:

使用所述语言选择模型处理所述训练示例的所述相对应的训练示例输入的所述相对应特征,以生成所述N种不同语言中的每一种语言的相对应预测概率,

基于所生成的相对应预测概率和相对应标记概率量度来生成相对应元组损失,以及

使用所生成的相对应元组损失更新所述语言选择模型的权重。

2.根据权利要求1所述的方法,其中,基于所生成的预测概率和所述相对应标记概率量度来生成所述相对应元组损失包括:

生成针对所述训练示例中的给定训练示例的所述元组损失中的给定元组损失,其中,生成所述给定元组损失包括:

基于所述给定训练示例的相对应标记概率量度与所述给定训练示例的相对应预测概率的比较,来确定各自针对小于N的相对应元组大小的一个或多个个体元组损失,其中,所述一个或多个个体元组损失至少包括针对相对应元组大小为2的成对损失;并且

基于所述一个或多个个体元组损失生成所述给定元组损失。

3.根据权利要求2所述的方法,其中,生成所述给定元组包括仅使用所述成对损失作为所述给定元组损失。

4.根据权利要求2所述的方法,其中,所述一个或多个个体元组损失进一步至少包括针对相对应元组大小为3的三个一组的损失,以及针对相对应元组大小为4的四个一组的损失。

5.根据权利要求4所述的方法,其中,生成所述给定元组损失基于至少所述成对损失、所述三个一组的损失以及所述四个一组的损失的加权组合。

6.根据权利要求5所述的方法,其中,所述成对损失在所述加权组合中的权重基于所测量的概率,所测量的概率指示仅指明用于话音处理的两种候选语言的用户的百分比。

7.根据权利要求1所述的方法,其中,使用所生成的相对应元组损失更新所述语言选择模型的所述权重包括:

跨所述语言选择模型反向传播所述元组损失。

8.根据权利要求1所述的方法,进一步包括继训练所述语言选择模型之后:

经由计算设备的至少一个麦克风接收捕获来自用户的当前口头话语的当前音频数据;

提取所述当前音频数据的一个或多个特征;

使用所述语言选择模型处理所述当前音频数据的所述一个或多个特征，以生成所述N种不同语言中的每一种语言的当前预测概率；

基于所述当前预测概率选择所述N种不用语言中的当前口头语言；以及

基于所选择的当前口头语言执行所述音频数据的话音至文本处理。

9.根据权利要求8所述的方法，其中，基于所选择的当前口头语言执行所述当前音频数据的话音至文本处理包括：

从多个候选话音辨识模型中选择与所选择的当前口头语言相对应的特定话音辨识模型；以及

使用所选择的话音辨识模型处理所述当前音频数据的所述特征以确定与所述当前口头话语相对应的一个或多个单词。

10.根据权利要求9所述的方法，进一步包括：

生成响应于所述一个或多个单词的内容；以及

提供所述内容以由所述计算设备渲染。

11.根据权利要求1所述的方法，进一步包括继训练所述语言选择模型之后：

经由计算设备的至少一个麦克风接收捕获来自用户的当前口头话语的当前音频数据；

基于第一候选口头语言和第二候选口头语言被指定为在所述计算设备处被利用和/或被所述计算设备的用户利用的语言，来识别所述N种不同语言中的至少所述第一候选口头语言和所述第二候选口头语言；

基于识别所述第一候选口头语言和所述第二候选口头语言：

使用针对所述第一候选口头语言的第一话音辨识模型发起所述音频数据的第一话音至文本处理，以及

使用针对所述第二候选口头语言的第二话音辨识模型发起所述音频数据的第二话音至文本处理；

提取所述当前音频数据的一个或多个特征；

与所述第一话音至文本处理和所述第二话音至文本处理同时地：

使用经训练的语言选择模型处理所述音频数据的所述一个或多个特征以生成所述N种不同语言中的每一种语言的当前预测概率，以及

基于所述当前预测概率确定所述当前口头话语是所述第一候选口头语言；

基于确定所述当前口头话语是所述第一候选口头语言：

在生成响应于所述当前口头话语的内容时使用在所述第一话音至文本处理期间所生成的输出。

12.根据权利要求11所述的方法，其中，基于所述当前预测概率确定所述当前口头话语是所述第一候选口头语言在所述第一话音至文本处理和所述第二话音至文本处理完成之前发生，并且进一步包括：

响应于确定所述当前口头话语是所述第一候选口头语言：

在所述第二话音至文本处理完成之前停止所述话音至文本处理，同时使得所述第一话音至文本处理完成。

13.根据权利要求11所述的方法，其中，确定所述当前口头话语是所述第一候选口头语言进一步基于：

在所述第一话音至文本处理期间所生成的输出的第一置信度测度，所述输出的所述第一置信度测度在所述第一话音至文本处理期间生成；以及

在所述第二话音至文本处理期间所生成的第二输出的第二置信度测度，所述第二输出的所述第二置信度测度在所述第二话音至文本处理期间生成。

14.根据权利要求1所述的方法，进一步包括：继训练所述语言选择模型之后：

经由计算设备的至少一个麦克风接收捕获当前口头话语的当前音频数据；

确定所述当前口头话语来自于所述计算设备的多个候选用户中的特定用户；

基于N种不同语言的子集被指定为由所述特定用户利用的语言来识别所述子集；

提取所述当前音频数据的一个或多个特征；

使用所述经训练的语言选择模型处理所述当前音频数据的所述一个或多个特征，以生成所述N种不同语言中的每一种语言的当前预测概率；以及

基于所述当前预测概率从所述子集中选择当前口头语言，其中，所述选择是响应于基于所述子集被指定为由所述特定用户利用的语言来识别所述子集而从所述子集进行的。

15.一种方法，包括：

经由计算设备的至少一个麦克风接收捕获来自用户的当前口头话语的当前音频数据；

提取捕获所述当前口头话语的所述当前音频数据的一个或多个特征；

使用语言选择模型来处理所述当前音频数据的所述一个或多个特征，以生成N种不同语言中的每一种语言的当前预测概率，所述语言选择模型使用元组损失进行训练；

基于与所述当前音频数据、所述计算设备和/或所述用户相关联的数据识别所述口头话语的M种候选语言，其中，所述M种候选语言包括两种或更多种语言，并且是所述N种不同语言的子集；

从所述M种候选语言中选择当前口头语言，其中，选择所述当前口头语言是基于所述M种候选语言的当前预测概率的比较；以及

基于所选择的当前口头语言执行所述音频数据的话音至文本处理。

16.根据权利要求15所述的方法，进一步包括：

在随所述当前音频数据的传输中接收所述M种候选语言的指示，其中，识别所述M种候选语言基于与所述当前音频数据相关联的所述数据，并且其中，所述数据包括在随所述当前音频数据的所述传输中所接收到的所述M种候选语言的所述指示。

17.一种系统，包括存储指令的存储器和一个或多个处理器，所述一个或多个处理器可操作以执行所述指令以使所述处理器执行根据前述权利要求中的任一项所述的方法。

# 训练和/或使用语言选择模型以自动确定用于口头话语的话音辨识的语言

## 背景技术

[0001] 人类可以参与与交互式软件应用的人机对话，该交互式软件应用在本文被称作“自动化助理”(也被称作“数字代理”、“聊天机器人”、“交互式个人助理”、“智能个人助理”、“助理应用”、“对话代理”等)。例如，人类(当他们与自动化助理交互式可以被称作“用户”)可以使用口头自然语言输入(即，话语)和/或通过提供文本(例如，键入的)自然语言输入向自动化助理提供命令和/或请求，该口头自然语言输入在一些情况下可以被转换为文本并且然后被处理。自动化助理通过提供响应性用户接口输出而对请求作出响应，该响应性用户接口输出可以包括可听和/或视觉用户接口输出。

[0002] 如上文所提到的，自动化助理可以将对应于用户的口头话语的音频数据转换为相对应的文本(或其它语义表示)。例如，音频数据可以基于经由客户端设备的一个或多个麦克风对用户的口头话语的检测而生成，该客户端设备包括用于使得用户能够与自动化助理交互的助理接口。自动化助理可以包括话音辨识引擎，该话音辨识引擎利用话音辨识模型来辨识在音频数据中所捕获的口头话语的各种特性，诸如该口头话语所产生的声音(例如，音素)、所产生的声音的顺序、话音的节奏、声调等。另外，话音辨识引擎可以识别这样的特性所表示的文本单词或短语。该文本然后可以由自动化助理在确定口头话语的响应内容时进一步处理(例如，使用自然语言理解(NLU)引擎和/或对话状态引擎)。话音辨识引擎可以由客户端设备和/或远离客户端设备但是与客户端设备网络通信的一个或多个自动化助理组件来实施。

[0003] 然而，许多话音辨识引擎被配置为仅辨识单一语言的话音。对于多语言用户和/或家庭来说，这样的单一语言话音辨识引擎可能无法令人满意，并且在以并非话音辨识引擎所支持的单一语言的附加语言接收到口头话语时可能导致自动化助理故障和/或提供错误的输出。这可以致使自动化助理不可用和/或引起计算和/或网络资源的过度使用。计算和/或网络资源的过度使用可能是由于用户在自动化助理故障或提供错误输出时需要提供以所支持的单一语言的另外的口头话语。这样的另外的口头话语必须由相对应的客户端设备和/或远程自动化助理组件附加处理，由此导致各种资源的附加使用。

[0004] 其它的话音辨识引擎可以被配置为辨识多种语言的话音，但是要求用户明确指定在给定时间应当在话音辨识中利用多种语言中的哪一种。例如，其它话音辨识引擎中的一些话音辨识引擎可能要求用户手动地指定在特定客户端设备处接收到的所有口头话语的话音辨识中要利用的默认语言。为了将该默认语言改变为另一种语言，可以要求用户与图形和/或可听接口进行交互以明确地更改默认语言。这样的交互可能引起接口的渲染、对用户经由接口所提供的输入的处理等中的计算和/或网络资源的过度使用。另外，可能经常出现用户在提供当前不是默认语言的口头话语之前忘记改变默认语言的情形。如上文所描述的，这可能致使自动化助理不可用和/或引起计算和/或网络资源的过度使用。

## 发明内容

[0005] 本文所描述的实施方式涉及用于训练和/或使用语言选择模型(其是神经网络模型或其它机器学习模型)以自动确定在音频数据中捕获的口头话语的特定语言。可以使用经训练的语言选择模型处理该音频数据的特征以生成N种不同语言中的每一种语言的预测概率,并且基于所生成的概率选择的特定语言。可以响应于选择了特定语言而利用针对该特定语言的话音辨识结果。许多实施方式涉及利用元组损失代替传统的交叉熵损失来训练语言选择模型。利用元组损失训练语言选择模型可以导致更加有效的训练,由此导致在训练期间利用更少的资源(例如,在训练期间处理训练示例时所利用的处理器和/或存储器资源)。附加地或可替选地,利用元组损失训练语言选择模型可以导致更加准确和/或鲁棒的模型——由此缓解了针对口头话语的错误语言选择。

[0006] 如本文所使用的,多个话音辨识模型可以被访问而用于话音辨识,并且话音辨识模型中的每种话音辨识模型可以被配置用于N种所支持话音辨识语言中的相对应语言。例如,第一话音辨识模型可以被配置用于在基于处理包括英语口头话语的音频数据来生成英语文本时使用,第二话音辨识模型可以被配置用于在基于处理包括法语口头话语的音频数据来生成法语文本时使用,第三可话音辨识模型以被配置用于在基于处理包括西班牙语口头话语的音频数据来生成西班牙语文本时使用。如上文所描述的,口头话语的特定语言可以至少部分地基于使用经训练的语言选择模型对捕获该口头话语的至少一部分的音频数据的处理而被选择。另外,针对特定语言的话音辨识结果可以响应于选择该特定语言而被利用。例如,可以仅利用符合特定语言的话音辨识模型来执行话音辨识,或者可以利用多个话音辨识模型,以及使用模型中的基于其符合特定语言而被利用的特定一种模型所生成的话音辨识结果来执行话音辨识。

[0007] 本文所公开的各种实施方式采用了大多数多语言用户仅说来自所支持的话音辨识语言的集合N的有限数量的语言这一观察。那些实施方式可以针对捕获口头话语的音频数据识别两种或更多种的候选语言M,并且基于仅比较所生成的该M种候选语言的概率来选择该口头话语的特定语言。换句话说,虽然利用经训练的语言选择模型来处理音频数据的至少一部分并且生成N种单独语言的概率,但是特定语言的选择可以基于作为N种所支持的话音辨识语言的子集的M种语言的概率。如本文更详细描述的,在考虑到以上观察的情况下,还利用在训练语言选择模型时所利用的元组损失。进一步地,利用元组损失代替仅交叉熵损失训练的语言选择模型可以导致N种所支持的话音辨识语言的概率的生成,而所述概率的生成在仅考虑那些语言中的M种时更可能引起正确的语言的选择。

[0008] 在其中针对给定音频数据仅考虑M种语言的实施方式中,该M种语言可以基于例如该M种语言在随音频数据的传输中被提供的指示(例如,该M种语言由客户端随该音频数据一起传输的指示),基于该M种语言关联于与该音频数据相关联的用户简档或其它标识符而被存储,和/或基于该M种语言关联于生成该音频数据的客户端设备而被存储。用于用户简档和/或设备的语言例如可以由用户手动地指定和/或基于用户对语言的过往使用(例如,跨一个或多个平台)、语言在客户端设备上过往使用等被自动指定。

[0009] 在一些实施方式中,语言选择模型可以是判别式N类分类器、长短期记忆(LSTM)网络,或者其它神经网络模型。可以使用诸如支撑向量机(SVM)模型的其它类型的模型。在其中采用SVM模型的一些实施方式中,元组损失可以与线性内核一起被应用,因为线性内核是

用原始形式的梯度下降算法被求解的。进一步地，使用监督或无监督学习以及利用本文所描述的元组损失对语言选择模型进行训练。出于简明的原因，关于监督学习描述了训练本文所描述的语言选择模型的许多实施方式。

[0010] 作为基于元组损失训练语言选择模型的一个特定示例，可以针对所支持的话音辨识语言N的集合中的每一种语言生成训练示例。每一个训练示例可以包括：对应于给定训练口头话语的音频数据的一个或多个特征的训练示例输入；和针对全体可能语言的集合N中的每一种语言的标记概率量度的训练示例输出。例如，针对每一个训练示例，可能存在针对所支持的话音辨识语言N的集合中的特定语言的正概率量度(例如，“1”和/或正概率量度的其它指示)，以及针对来自全体可能语言的集合N的其它每种语言的负概率量度(例如，“0”和/或负概率量度的其它指示)。可以基于训练示例来训练语言选择模型，其中，元组损失是基于训练示例生成的，并且元组损失被用于更新语言选择模型的权重(例如，通过反向传播)。

[0011] 每一个元组损失可以被生成为一个或多个个体元组损失的函数，该一个或多个个体元组损失各自针对小于N(语言选择模型所预测的概率的数量)的相对应元组大小。例如，给定训练模型的元组损失可以至少部分地基于成对损失而生成，其中成对损失是针对大小2的元组的，并且是基于将N个概率中的所有对的预测概率(其中预测概率是通过使用语言选择模型处理给定训练示例的训练示例输入生成的)与如训练示例输出所指示的所有对的标记概率量度的标记概率相比较生成的。

[0012] 在一些实施方式中，成对损失(L(y,z))可以由以下等式1所表示：

[0013]
$$L(y,\boldsymbol{z}) = -\mathbf{E}_{k\neq y}\left[log\frac{\exp(\boldsymbol{z}_y)}{\exp(\boldsymbol{z}_y)+\exp(\boldsymbol{z}_k)}\right]$$

$$= \mathbf{E}_{k\neq y}\left[log\left(\exp(\boldsymbol{z}_y)+\exp(\boldsymbol{z}_k)\right)\right] - \boldsymbol{z}_y \quad (等式1)$$

[0014] 其中$z_k$是(N种语言中的)第k种语言的预测非归一化概率，其中$E_s$表示集合s的期望。子集$S_y^n$可以是包括具有n个元素的正确标记y的所有元组集，其中1<n≤N，其中子集$S_y^n$中的元组的数量可以对应于M种语言。如从以上描述所理解的，不同于交叉熵损失，成对损失不使得正确标记的概率最大化，而使得所有其它标记的概率相等地最小化。相反，利用成对损失，使得所有其它(不正确)标记的概率的最小化是不相等。同样，这在考虑到大多数多语言用户可能仅说来自所支持话音辨识语言的集合N的有限数量的语言这一观察——以及考虑到在利用成对损失(以及可选地其它的个体元组损失)对模型训练之后在推导时间仅考虑语言的子集的语言选择情形中可能是有利的。

[0015] 虽然上文描述了成对损失，但是在各种实施方式中，元组损失进一步是附加的个体元组损失的函数。例如，针对给定训练示例的元组损失可以至少部分地基于三个一组(tri-wise)的损失(针对元组大小为3)，至少部分基于四个一组(four-wise)的损失(针对元组大小4)等进一步生成。在一些实施方式中，以上等式1中的成对损失可以被一般化表示为以下的等式2以用于确定从n个标记的子集产生标记的损失：

[0016]
$$L^n(y,\boldsymbol{z}) = \mathbf{E}_{S_y^n}\left[log\sum_{k\in S_y^n}\exp(\boldsymbol{z}_k)\right] - \boldsymbol{z}_y \quad (等式2)$$

[0017] 其中$\mathbf{S}_y^n$是$S_y$中大小为n的所有元组。因此，$\mathbf{S}_y^n$中存在元组的$\|\mathbf{S}^n\| = \binom{N-1}{n-1}$种组合。例如，如果N＝50种不同的口头语言并且针对成对损失n＝2，则语言选择模型可以针对$\mathbf{S}_y^n$中的元组的组合中的每一种组合或者针对1176个元组确定个体元组损失。作为另一个示例，如果N＝50种不同的口头语言并且针对三个一组的损失n＝3，则语言选择模型可以针对$\mathbf{S}_y^n$中的元组的组合中的每一种组合或者针对18424个元组确定个体元组损失。

[0018] 基于紧接在前的等式2，总元组损失L(y,z)可以基于每种语言的预测概率而被确定为所有1<n≤N的不同大小的所有个体元组损失的加权和。在一些实施方式中，所有个体元组损失的加权和可以由以下的等式3所定义：

[0019] $$L(y,\boldsymbol{z}) = \mathbf{E}_{\mathbf{S}^n\sim D}[L^n(y,\boldsymbol{z})] = \sum_{n=2}^{N} p_n L^n(y,\boldsymbol{z}) \quad \text{（等式3）}$$

[0020] 其中$p_n$是大小为n的元组的概率并且$L^n(y,z)$是与$p_n$相关联的损失。大小为n的元组的概率$p_n$可以对应于与大小为n的语言的量相关联的多语言用户、设备和/或请求的百分比。例如，如果90％的多语言用户仅指明了两种预定义语言，则$p_2$可以为0.9。作为另一个示例，如果7％的用户指明了三种预定义语言，则$p_3$可以为0.07。因此，$p_n$有效地使得总元组损失朝向更可能发生的元组大小的个体元组损失偏移(例如，最大程度地偏向成对损失，随后是三个一组的损失等)。如上文所理解的，在一些实施方式中，除了大小小于N的元组的个体元组损失之外，元组损失还部分基于大小为N的元组的个体元组损失。这可以被视为其中n等于N的个体元组损失的特殊情况，并且等同于交叉熵损失。然而，在那些实施方式的许多实施方式中，交叉熵损失的权重$p_N$可以基于例如极少用户实际上将所有所支持的语言N都指定为口头话语的候选语言而为最小。因此，虽然总元组损失可以是交叉熵损失的函数，但是其也是小于N的元组的个体元组损失的函数——并且这种小于N的元组的元组损失可以显著地比交叉熵损失更重的共同加权。

[0021] 提供以上描述作为本公开一些实施方式的概述。在下文更详细地描述那些实施方式以及其它实施方式的进一步描述。

[0022] 在一些实施方式中，阐述了一种由一个或多个处理器实现的方法，并且该方法包括生成多个训练示例。生成训练示例中的每一个训练示例基于捕获相对应的人类话语的相对应音频数据，以及指示相对应人类话语的相对应的口头语言的相对应的标记。相对应的口头语言是所要辨识的N种不同语言中的一种语言，其中N是大于10的整数。进一步地，每一个训练示例包括相对应训练示例输入以及相对应的训练示例输出，该相对应的训练示例输入包括相对应音频数据的相对应特征，该相对应的训练示例输出包括针对所要辨识的N种不同语言的中的每一种语言的相对应标记概率量度。该相对应标记概率量度基于相对应标记，包括对应于相对应口头语言的相对应正概率量度标记，以及针对所有其它相对应标记概率量度的相对应负概率量度标记。该方法进一步包括基于训练示例训练语言选择模型。训练语言选择模型包括使用语言选择模型处理训练示例的相对应训练示例输入的相对应特征以生成N种不同语言中的每一种语言的相应预测概率，基于所生成的相对应预测概率和相对应标记概率量度来生成相应元组损失，以及使用所生成的相对应元组损失来更新语言选择模型的权重。

[0023] 在一些实施方式中，基于所生成的预测概率和相对应标记概率量度生成相对应元

组损失包括针对训练示例中的给定训练示例生成元组损失中的给定元组损失。在一些实施方式中，生成给定元组损失进一步包括基于给定训练示例的相对应标记概率量度与给定训练示例的相对应预测概率的比较来确定各自针对与小于N的相对应元组大小的一个或多个个体元组损失。一个或多个个体元组损失至少包括针对相对应元组大小为2的成对损失。在一些实施方式中，生成给定元组损失进一步包括基于一个或多个个体元组损失生成给定元组损失。

[0024] 在一些实施方式中，生成给定元组包括仅使用成对损失作为给定元组损失。在其它实施方式中，一个或多个个体元组损失进一步至少包括针对相对应元组大小为3的三个一组的损失，以及针对相对应元组大小为4的四个一组的损失。在一些其它实施方式中，生成给定元组损失基于至少成对损失、三个一组的损失、四个一组的损失的加权组合。在一些其它实施方式中，成对损失在加权组合中的权重基于所测量的概率，该所测量的概率指示仅指明用于话音处理的两种候选语言的用户的百分比。

[0025] 在一些实施方式中，使用所生成的相对应元组损失更新语言选择模型的权重包括跨语言选择模型反向传播元组损失。

[0026] 在一些实施方式中，继训练该语言选择模型之后，该方法可以进一步包括经由计算设备的至少一个麦克风接收捕获来自用户的当前口头话语的音频数据，提取当前口头话语的一个或多个特征，使用语言选择模型处理当前口头话语的一个或多个特征以生成N种不同语言中的每一种语言的当前预测概率，基于当前预测概率选择N种不用语言中的当前口头语言，基于所选择的当前口头语言执行音频数据的话音至文本处理。在那些实施方式的一些实施方式中，该方法可以进一步包括从多个候选话音辨识模型中选择与所选择的当前口头语言相对应的特定话音辨识模型，并且使用所选择的话音辨识模型处理音频数据的特征以确定与当前口头话语相对应的一个或多个单词。在那些实施方式的一些实施方式中，该方法可以进一步包括生成响应于该一个或多个单词的内容，并且提供该内容以便由该计算设备所渲染。

[0027] 在一些实施方式中，该方法可以进一步包括继训练语言选择模型之后，经由计算设备的至少一个麦克风接收捕获来自用户的当前口头话语的当前音频数据。在一些实施方式中，该方法可以进一步包括基于第一候选口头语言和第二候选口头语言被指定为在计算设备处被利用和/或被该计算设备的用户所利用的语言，来识别该N种不同语言中的至少第一候选口头语言和第二候选口头语言。在一些实施方式中，该方法可以进一步包括基于识别第一候选口头语言和第二候选口头语言，使用针对第一候选口头语言的第一话音辨识模型发起音频数据的第一话音至文本处理，并且使用针对第二候选口头语言的第二话音辨识模型发起音频数据的第二话音至文本处理。在一些实施方式中，该方法可以进一步包括提取当前音频数据的一个或多个特征。在一些实施方式中，该方法可以进一步包括与第一话音至文本处理和第二话音至文本处理同时地，使用经训练的语言选择模型处理音频数据的一个或多个特征以生成N种不同语言中的每一种语言的当前预测概率，以及基于当前预测概率确定当前口头话语是第一候选口头语言。在一些实施方式中，该方法可以进一步包括基于确定当前口头话语是第一候选口头语言，在生成响应于当前口头话语的内容时使用在第一话音至文本处理期间所生成的输出。

[0028] 在那些实施方式的一些实施方式中，基于当前预测概率确定当前口头话语是第一

候选口头语言在第一话音至文本处理和第二话音至文本处理完成之前发生，并且进一步包括响应于确定当前口头话语是第一候选口头语言，在第二话音至文本处理完成之前停止该第二话音至文本处理，同时使得第一话音至文本处理得完成。

[0029] 在那些实施方式的一些那些实施方式中，确定当前口头话语是第一候选口头语言进一步基于在第一话音至文本处理期间所生成的输出的第一置信度测度，该输出的第一置信度测度在该第一话音至文本处理期间生成。在那些实施方式的一些中，确定当前口头话语是第一候选口头语言进一步基于在第二话音至文本处理期间所生成的第二输出的第二置信度测度，该第二输出的第二置信度测度在第二话音至文本处理期间生成。

[0030] 在一些实施方式中，该方法可以进一步包括继训练该语言选择模型之后经由计算设备的至少一个麦克风接收捕获当前口头话语的当前音频数据，确定该当前口头话语来自于计算设备的多个候选用户中的特定用户，基于N种不同语言的子集被指定为由该特定用户所利用的语言而识别该子集，以及提取当前音频数据的一个或多个特征。在一些实施方式中，该方法进一步包括继训练语言选择模型之后使用经训练的语言选择模型处理当前音频数据的一个或多个特征以生成N种不同语言中的每一种语言的当前预测概率，并且基于该当前预测概率从子集中选择当前口头语言，其中选择是响应于基于子集被指定为由特定用户所利用的语言来识别该子集而从该子集进行的。

[0031] 提供以上描述是作为本公开的一些实施方式的概述。以下更详细地描述那些实施方式以及其它实施方式的进一步描述。

[0032] 在一些实施方式中，阐述了一种由一个或多个处理器实现的方法，并且该方法包括经由计算设备的至少一个麦克风接收捕获来自用户的当前口头话语的当前音频数据，提取捕获当前口头话语的当前音频数据的一个或多个特征，使用使用元组损失训练的语言选择模型来处理当前口头话语的一个或多个特征以生成N种不同语言中的每一种语言的当前预测概率，基于当前预测概率选择N种不用语言中的当前口头语言，以及基于所选择的当前口头语言执行音频数据的话音至文本处理。

[0033] 其它实施方式可以包括存储指令的非暂时性计算机可读存储介质，该指令能够由一个或多个处理器(例如，中央处理单元(CPU)、图形处理单元(GPU)和/或张量处理器(TPU))执行以执行诸如在上文和/或在本文其它地方所描述的一种或多种方法的方法。再其它的实施方式可以包括一个或多个计算机的系统，该一个或多个计算机包括一个或多个处理器，该一个或多个处理器可操作以执行所存储的指令以执行诸如在上文和/或在本文其它地方所描述的方法中的一种或多种方法。

[0034] 应当理解的是，以上概念以及在本文更详细描述的附加概念的所有组合都被认为本文所公开主题的一部分。例如，出现在本公开的结尾处的所请求保护的主题的所有组合都被认为本文所公开主题的一部分。

## 附图说明

[0035] 图1示出了根据本文所公开的各种实施方式的利用经训练的语言选择模型来选择自动化助理与用户交互的语言的示例系统。

[0036] 图2示出了语言选择模型的一个示例，利用元组损失训练该语言选择模型的示例，以及一旦经过训练利用该语言选择模型的示例。

[0037] 图3是示出根据本文所公开的实施方式的用于训练语言选择模型的示例方法的流程图。

[0038] 图4是示出根据本文所公开的实施方式的用于使用经训练的语言选择模型来选择话音辨识语言的示例方法的流程图。

[0039] 图5是示出根据本文所公开的实施方式的用于使用经训练的语言选择模型来选择话音辨识语言的另一种示例方法的流程图。

[0040] 图6是根据本文所公开的实施方式的示例计算机系统的框图。

## 具体实施方式

[0041] 图1示出了根据本文所公开的各种实施方式的利用经训练的语言选择模型150来选择自动化助理104与用户130交互的语言的示例系统。自动化助理104可以部分经由在诸如客户端计算设备118(例如,便携式计算设备132)的一个或多个客户端设备处提供的自动化助理126并且部分经由诸如服务器设备102(例如,其可以形成经常被称为“云基础设施”或简称为“云”)的一个或多个远程计算设备112进行操作。当在本文使用“自动化助理104”时,其可以是指104和126中的一个或二者。用户130可以经由客户端计算设备118的助理接口128与自动化助理104进行交互。助理接口128包括用户接口输入设备和用户接口输出设备以由自动化助理126在与用户130交互时使用。

[0042] 助理接口128接受用户130的指向自动化助理104的用户接口输入,并且渲染来自自动化助理104的响应于该用户接口输入的内容以向用户140呈现。助理接口128可以包括麦克风、扬声器、显示面板、相机、触摸屏显示器中的一个或多个,和/或客户端计算设备118的任何其它用户接口设备。助理接口128还可以包括显示器、投影仪、扬声器,和/或可以被用来渲染来自自动化助理104的内容的客户端计算设备118的任何其它(多个)用户接口输出设备。用户可以通过向助理接口128提供言语、文本或图形输入来初始化自动化助理104以使自动化助理104执行功能(例如,提供数据、控制外围设备、访问代理等)。在一些实施方式中,客户端计算设备118可以包括显示设备,该显示设备可以是包括触摸接口的显示面板,该触摸接口用于接收触摸输入和/或手势以允许用户经由该触摸接口控制客户端计算设备的应用。在一些实施方式中,客户端计算设备118可能缺少显示设备,由此提供可听的用户接口输出,而并不提供图形用户界面输出。此外,客户端计算设备118可以提供用户接口输入设备,诸如麦克风,以用于接收来自用户130(以及来自附加的未示出的用户)的口头自然语言输入。

[0043] 客户端计算设备118可以通过诸如互联网的一个或多个网络114与远程计算设备112进行通信。客户端计算设备118可以将计算任务卸载至远程计算设备112,以便例如节省客户端设备118处的计算资源和/或采用远程计算设备112处可用的更加鲁棒的资源。例如,远程计算设备112可以托管自动化助理104,并且客户端计算设备118可以将在一个或多个助理接口接收到的输入传输至远程计算设备112。然而,在一些实施方式中,自动化助理104可以由客户端计算设备118处的自动化助理126托管。在各种实施方式中,自动化助理104的全部或少于全部的方面可以由客户端计算设备118处的自动化助理126来实现。在那些实施方式的一些实施方式中,自动化助理104的方面经由客户端计算设备118处的本地自动化助理126实现并且与实现自动化助理104的其它方面的远程计算设备112对接。

[0044] 远程计算设备112可选地可以经由用户简档为多个用户以及它们的相关联助理应用服务。在一些实施方式中，服务器设备102可以存储服务器用户简档120。在一些其它实施方式中，客户端计算设备118可以存储客户端用户简档122。在自动化助理104的全部或少于全部的方面经由客户端计算设备118的本地自动化助理126实现的实施方式中，本地自动化助理126可以是与客户端设备118的操作系统分离的应用(例如，安装在操作系统的“顶端”)——或者可以可替选地，可以直接由客户端设备118的操作系统实现(例如，被认为是操作系统的应用但是与操作系统整合)。

[0045] 在一些实施方式中，服务器设备102可以包括语言选择模型150和/或客户端计算设备118可以包括语言选择模型160。语言选择模型150和语言选择模型160可以是相同的模型，或者语言选择模型160可选地可以被优化用于在更加资源受限的客户端计算设备118上的使用的变体。而且，在各种实施方式中，语言选择模型150可以在服务器设备102上实现而无需在客户端计算设备118上实现语言选择模型160，或者语言选择模型160可以在客户端计算设备118上实现而无需在服务器设备102上实现语言选择模型150。

[0046] 如本文所描述的，自动化助理104可以在选择对应于所接收的口头人类话语的特定语言时利用语言选择模型150和/或自动化助理126可以在选择对应于所接收的口头人类话语的特定语言时利用语言选择模型160。例如，自动化助理104可以使用语言选择模型150处理所接收的音频数据的至少一部分以生成N种所支持的话音辨识语言中的每一种话音辨识语言的概率。进一步地，自动化助理104可以利用所生成的概率来选择那N种所支持语言中的一种语言作为音频数据所捕获的口头话语的特定语言。例如，自动化助理104可以将第一和第二语言识别为口头话语的候选语言，并且至少部分基于(来自N个概率的)第一语言的第一概率与(来自N个概率的)第二语言的第二概率的比较来选择第一语言或第二语言。注意到，在各种实施方式中，自动化助理104和/或自动化助理126也可以在选择特定语言时依赖于一个或多个附加信号，诸如本文所描述的其它信号。

[0047] 在一些实施方式中，利用所选择的特定语言来仅选择话音辨识模型136中的相应的一个话音辨识模型以执行音频数据的话音至文本(STT)处理。在一些实施方式中，STT处理可能已经由多个话音辨识模型136与使用语言选择模型150的处理并行地处执行了。例如，可以在使用语言选择模型150执行处理的同时针对M种候选语言中的每一种候选语言初始化STT处理。但是，在那些实施方式的一些实施方式中，所选择的特定语言用来选择仅由话音辨识模型136中的相应的一个话音辨识模型所生成的输出，以及可选地停止使用不对应于所选择的特定语言的话音辨识模型136的处理。以下关于图2和3更详细地描述用于训练语言选择模型150的示例方法。

[0048] 在一些实施方式中，远程计算设备112可以包括话音辨识引擎134，该话音辨识引擎134可以处理在助理接口128处所接收的音频数据以确定该音频数据中所体现的口头话语的文本和/或其它语义表示。话音辨识引擎134可以在确定音频数据中所体现的口头话语的文本和/或其它语义表示时可以采用一个或多个话音辨识模型136。如本文所描述的，可以提供多个话音辨识模型136，并且每一个话音辨识模型可以针对相对应的语言。例如，第一话音辨识模型可以针对英语，第二话音辨识模型可以针对法语，第三话音辨识模型针对西班牙语，第四话音辨识模型针对汉语，第五话音辨识模型针对日语等。

[0049] 在一些实施方式中，话音辨识模型136各自包括用于确定对应于音频数据中所体

现的口头话语的文本(或其它语义表示)的一个或多个机器学习模型和/或统计模型。在一些实施方式中,话音辨识引擎134可以利用话音辨识模型136中的一个话音辨识模型136确定音频数据中所包括的针对相对应语言的音素,并且然后基于所确定的音素生成针对该相对应语言的文本。在一些实施方式中,话音辨识模型接收例如数字音频数据的形式的语音输入的音频录制,并且将该数字音频数据转换为一个或多个文本符号(例如,STT处理)。这样的功能所使用的一个或多个模型总体上对音频信号和语言中的音素单元之间的关系连同该语言中的单词序列一起进行建模。在一些实施方式中,话音辨识模型可以是声学模型、语言模型、发音模型等,以及对这样的模型中的一种或多种这样的模型的组合功能进行建模。在一些实施方式中,例如,话音辨识模型可以被实现为包括多条路径或通路的有限状态解码图。

[0050] 进一步地,如本文所描述的,在确定多种话音辨识模型136中的哪些应当在处理音频数据以生成语义和/或文本表示时被加以利用时和/或在选择应当利用哪些语义和/或文本表示时可以利用附加的语言选择模型150。例如,在那些实施方式的一些实施方式中,语言选择模型150用来生成口头话语对应于N种不同语言中的每一种语言的预测概率,其中多个话音辨识模型136对应于该N种不同语言中的每一种语言。给定语言的预测概率中的每一个预测概率可以构成有关用户说什么语言的“猜测”或“预测”。

[0051] 当用户130与客户端计算设备118处的自动化助理126通信时,用户130可以向客户端计算设备118的助理接口128提供口头自然语言输入。该口头自然语言输入可以被转换为音频数据,该音频数据可以被客户端语言模型124所处理,该客户端语言模型124诸如用于识别音频数据是否体现了用于调用自动化助理126的调用短语的调用短语模型。在一些实施方式中,调用短语模型可以在客户端计算设备118处用来确定用户130是否想要调用自动化助理104。当用户向助理接口128提供了自然语言输入,并且该自然语言输入包括用于调用自动化助理104的调用短语时,客户端计算设备118可以使得服务器设备102处的自动化助理104接收该自然语言输入和/或来自用户130的后续自然语言输入。

[0052] 例如,响应于确定用户130想要调用客户端计算设备118处的自动化助理104,可以在客户端计算设备118和服务器设备102之间建立一个或多个通信信道。其后,随着用户继续向助理接口128提供自然语言输入,该自然语言输入将被转换为然后通过网络114传输并且由服务器设备102处理的数据。该自然语言输入可以由服务器设备102使用语言选择模型150处理以生成该自然语言输入对应于N种不同语言中的每一种语言的预测概率。基于预测概率,一个或多个话音辨识模型136可以被选择作为针对每一个自然语言输入的适当模型。

[0053] 在一些实施方式中,一个或多个话音辨识模型136中仅对应于特定口头语言的一个话音辨识模型136可以被选择用于自然语言输入的STT处理。在一些其它实施方式中,自然语言输入的STT处理可能已经由一个或多个话音辨识模型136中对应于特定口头语言以及N种不同语言中的至少一种附加语言的多个话音辨识模型136与使用语言选择模型150的处理并行地执行。例如,可以在使用语言选择模型150执行处理的同时针对M种候选语言中的每一种候选语言初始化STT处理。但是,在那些实施方式的一些实施方式中,所选择的特定语言用来选择仅由话音辨识模型136中的相对应的一个话音辨识模型136生成的输出,并且可选地基于排名停止使用不对应于所选择的特定语言的话音辨识模型136的处理。

[0054] 图2示出了语言选择模型的一个示例(图2的示例中的LSTM模型250),利用元组损

失训练该语言选择模型250的示例，以及一旦经过训练利用该语言选择模型250的示例。训练示例280可以存储在一个或多个数据库中，训练示例280中的每一个训练示例280对应于相对应口头语言的口头人类话语。进一步地，训练示例280中的每一个训练示例280可以被声学模型220处理以针对训练示例280中的每一个训练示例280提取音频数据260的一个或多个特征——被表示为特征序列x，以及标记概率量度236——被表示为标记y，其中$y \in \{1,\cdots,N\}$，并且其中N是全部可能语言的集合，这表明目标语言来自于全部可能语言的集合N。音频数据260的一个或多个特征可以用作针对语言选择模型的训练示例输入，该语言选择模型诸如图1中的语言选择模型150、160，其在图2中被表示为长短期存储器(LSTM)模型250(但是如本文所描述的，可以利用其它网络架构)。标记概率量度236指示针对对应于给定训练示例的语言的正概率量度，并且指示针对所有其它语言的负概率量度。

[0055] 在一些实施方式中，在使用LSTM模型250处理音频数据260的一个或多个特征之前，该音频数据的一个或多个特征可以穿过级联层。该级联层可以允许在推导时间利用滑动窗口方法，这在本文更详细的描述(例如，参考图4-6)。通过使用级联层，LSTM模型250生成的输出可以更大，但是作为使用级联层的结果，训练明显更快且LSTM250更加鲁棒。例如，级联层可以级联音频数据的相邻分段使得输入数量减半。

[0056] 在一些实施方式中，在LSTM模型250的每一个层之后，可以提供投影层以减少用于LSTM模型250的参数的大小。通过增加投影层并且减小用于LSTM模型250的参数的大小，LSTM模型250的训练以及使用LSTM模型250的推导可以明显加速训练和推导而并不损害性能。在一些实施方式中，在LSTM模型250之后，可以提供时间池化层(temporal pooling layer)以将LSTM模型250的最后输出映射至全部可能语言的集合N中的每一种语言的线性投影。通过增加池化层，经训练的神经网络以最小延时来执行且并不要求任何上下文或填充。

[0057] 继续参考图2，在一些实施方式中，z可以是语言选择模型的最后一层的N维输出，并且$z=f(x;w)$可以表示特征序列x在N种不同语言上的非归一化分布，其中w可以表示语言选择模型的参数。在一些实施方式中，$z_k$可以是N种不同语言中的第k种语言的预测非归一化概率。语言选择模型可以被训练以针对全部可能语言的集合N中的每一种语言输出概率，并且可以从来自全部可能语言的集合N的子集S选择语言。子集S可以利用本文所描述的技术来识别。例如，可以基于所接收的音频数据与用户简档相关联并且该用户简档将子集S指示为与该用户简档相关联的说话者所说的语言而针对所接收的音频数据选择子集S。

[0058] 在一些实施方式中，由元组损失引擎240针对每一个训练示例生成元组损失包括将针对每一个训练示例的标记概率量度236与针对每一个训练示例的预测概率238相比较，并且确定关于n个元组的加权组合。标记概率量度236可以是指示一个或多个值的向量，该一个或多个值指示针对给定训练示例，全部口头语言的集合N中的哪一种口头语言应当通过音频数据206的一个或多个特征而被辨识向量。在一些实施方式中，标记概率量度236可以包括针对由给定训练示例的音频数据260的一个或多个特征捕获的口头语言的正概率量度(例如，值“1”)，以及针对来自全部口头语言的集合N中的所有其它口头语言的负概率量度(例如，值“0”)。例如，假设在训练期间已经利用语言选择模型来生成针对给定训练示例的音频数据的预测概率238[0.7,0.3,0.0,…,0.0]，并且训练示例具有标记概率量度236[1,0,0,…,0]。在这样的示例中，可以通过将预测概率238[0.7,0.3,0.0,…,0.0]的一个或

多个大小为n的元组中的全部(例如,“0.7”和“0.3”,“0.7”和“0.0”,“0.3”和“0.0”等)与标记概率量度236[1,0,0,…,0]的一个或多个大小为n的元组中的全部(例如,“1”和“0”,“0”和“0”等)相比较来生成总元组损失。

[0059] 在一些实施方式中,元组损失至少部分地基于成对损失,但是这并非意在作为限制。如本文(例如,关于发明内容)所阐述的成对损失可以由以下等式1所表示:

[0060]
$$L(y,\mathbf{z}) = -\mathbf{E}_{k\neq y}\left[log\frac{\exp(\mathbf{z}_y)}{\exp(\mathbf{z}_y)+\exp(\mathbf{z}_k)}\right]$$
$$= \mathbf{E}_{k\neq y}\left[log\left(\exp(\mathbf{z}_y)+\exp(\mathbf{z}_k)\right)\right]-\mathbf{z}_y \quad (等式\ 1)$$

[0061] 其中$E_s$表示集合s的期望值。子集$S_y^n$可以是包括具有n个元素的正确标记y的所有元组集,其中1<n≤N。等式1的成对损失可以被一般化为等式2以用于确定从n个标记的子集产生标记的损失:

[0062]
$$L^n(y,\mathbf{z}) = \mathbf{E}_{S_y^n}\left[log\sum_{k\in S_y^n}\exp(\mathbf{z}_k)\right]-\mathbf{z}_y \quad (等式2)$$

[0063] 其中$\mathbf{S}_y^n$是$S_y$中大小为n的所有元素。因此,$\mathbf{S}_y^n$中存在$\|S^n\| = \binom{N-1}{n-1}$个元组。

[0064] 基于等式2,总损失L(y,z)可以基于每种语言的预测概率而被确定为所有1<n≤N的不同大小的所有元组损失的加权和。该所有元组损失的加权和由以下等式3中的元组损失函数所定义:

[0065]
$$L(y,\mathbf{z}) = \mathbf{E}_{S^n\sim D}[L^n(y,\mathbf{z})] = \sum_{n=2}^{N} p_n L^n(y,\mathbf{z}) \quad (等式3)$$

[0066] 其中$p_n$是大小为n的元组的概率并且$L^n$(y,z)是与$p_n$相关联的损失。大小为n的元组的概率$p_n$对应于在用户简档或附加用户简档中指定了预定义的n种语言的用户的百分比。例如,如果90%的用户指定了两种预定义语言,则$p_2$可以为0.9。作为另一个示例,如果7%的用户指定了三种预定义语言,则$p_3$可以为0.07。通过使用元组损失训练语言选择模型,该系统可以明显加速训练和推导而并不损害性能。

[0067] 在推导时,预测概率238均可以与针对N种不同语言中的每一种语言的相对应话音辨识模型$232_1$-$232_N$相关联。使用LSTM模型250的系统可以被配置为通过处理特征序列x以确定与来自全部可能语言的集合N的当前口头语言的当前口头话语相对应的一个或多个单词而基于该预测概率在话音辨识模型$232_1$-$232_N$之间进行选择。例如,使用LSTM模型250的系统可以接收与英语的当前口头话语的音频数据相对应的特征序列x。基于用户简档,可以获知提供口头人类话语的用户能够说英语和西班牙语。基于特征序列x,该系统可以确定该口头人类话语为英语的预测概率为0.7,并且该口头人类话语为西班牙语的预测概率为0.3。

[0068] 因此,使用LSTM模型250的系统可以基于与英语相关联的预测概率0.7大于与西班牙语相关联的预测概率0.3而选择与英语相关联的话音辨识模型,诸如第一辨识模型$232_1$,而不是与西班牙语相关联的话音辨识模型,诸如第二辨识模型$232_2$。如果用户能够说两种语言,则该两种语言具有成对关系(例如,英语和西班牙语、西班牙语和德语、德语和法语等之间的成对关系)。在一些实施方式中,一旦选择了话音辨识模型,就可以执行与特征序列x相关联的音频数据的STT处理以确定与口头人类话语相对应的一个或多个单词234。进一步

地，该系统可以生成响应于该一个或多个单词的内容236来生成内容，并且将该内容提供至计算设备以渲染该内容。在一些实施方式中，如本文所描述的，音频数据的STT处理可以与使用LSTM模型250针对口头话语选择语言并行地执行，并且在生成响应性内容时利用对应于所选择的语言的STT输出。

[0069] 作为另一个示例，在推导时，考虑第一预测概率分布[0.3,0.4,0.2,0.1]和第二预测概率分布[0.3,0.25,0.25,0.2]，其中该第一预测概率对应于作为口头话语的“正确”语言的第一语言，并且其中每种预测概率分布中的预测概率中的每一个预测概率对应于N种不同语言中的每一种语言。进一步地，第一预测概率分布可以对应于与使用交叉熵损失函数训练的语言选择模型相关联的预测概率分布，并且第二预测概率分布可以对应于与由元组损失引擎240使用元组损失函数——诸如等式(3)的损失函数——所训练的语言选择模型相关联的预测概率分布。基于第一预测概率分布，如由概率0.4所证明的，使用交叉熵损失函数训练的语言选择模型可以提供指示口头人类话语对应于第二口头语言的概率。然而，第二语言的该选择不正确。基于第二预测概率分布，如由概率0.3所证明的，使用元组损失函数训练的语言选择模型可以提供指示口头人类话语对应于第一口头语言的概率。因此，通过在训练期间使用元组损失函数，语言选择模型可以在推导时提供更加准确的结果，这减少了用户所接收的输入的数量，节省了计算资源，并且为用户提供了整体上更好的体验。

[0070] 图3是示出根据本文所公开的实施方式的用于训练语言选择模型的示例方法300的流程图。为了方便，该流程图的操作参考执行该操作的系统来描述。该系统可以包括各种计算机系统的各种组件，诸如图1中所描绘的一个或多个组件。此外，虽然方法300的操作是以特定顺序被示出，但是这并非意在作为限制。一个或多个操作可以被重新排序、被省略或者被添加。

[0071] 在框352，该系统基于捕获相对应人类话语的相对应音频数据以及指示该相对应人类话语的相对应口头语言的相对应标记来生成多个训练示例。每一个训练示例的相对应口头语言是要由该系统辨识的N种不同语言中相对应的一种语言。例如，该系统可以基于英语语言的人类话语来生成训练示例。人类话语可以与音频数据以及指示人类话语为英语语言的标记相关联。进一步地，框352可以包括一个或多个子框。

[0072] 在子框352A，该系统确定包括相对应音频数据的相对应特征的相对应训练示例输入。继续以上示例，该系统可以根据相对应音频数据确定音频数据的一个或多个特征，诸如梅尔频率倒谱系数(MFCC)、对数梅尔滤波器组(log-mel-filterbank)特征和/或其它特征。

[0073] 在子框352B，该系统确定相对应的训练示例输出，该相对应的训练示例输出包括针对要辨识的N种不同语言中的每一种语言的相对应标记概率量度。进一步地，相对应标记概率量度可以包括与相对应训练示例输入的相对应口头语言相对应的相对应正概率量度标记，以及针对所有其它的相对应标记概率量度的相对应负概率量度标记。继续以上示例，该系统可以根据英语语言的人类话语而确定标记概率量度，该标记概率量度可以被表示为向量，使得值“1”对应于英语语言的正概率量度的以及值“0”对应于N种不同语言中的所有其它语言的负概率量度的。

[0074] 在框354，该系统基于训练示例训练语言选择模型。继续以上示例，该系统可以接收训练示例，该训练示例包括捕获英语语言的人类话语的音频数据的特征的相对应训练示

例输入，以及本文所描述的标记概率量度的相对应的训练示例输出(例如，关于图2)。进一步地，框354可以包括一个或多个子框。

[0075] 在子框354A，该系统可以使用语言选择模型处理训练示例输入的相对应特征以生成口头语言中的每种口头语言的预测概率。继续以上示例，该系统可以处理英语的人类话语的特征，并且至少生成指示人类话语的特征很可能对应于英语语言的第一预测概率0.7，并且至少生成指示人类话语的特征不大可能——但是有可能——对应于西班牙语的第二预测概率0.2。同样可以生成其它语言的其它概率。

[0076] 在子框354B，该系统基于所生成的预测概率以及相对应标记概率量度生成元组损失。本文中描述了生成元组损失的示例，并且如所描述的，该元组损失可选地可以是诸如成对损失、三个一组的损失等的各种个体元组损失的函数。

[0077] 在子框354C，该系统使用所生成的元组损失更新语言选择模型的权重。继续以上示例，该系统可以通过跨语言选择模型反向传播元组损失来更新语言选择模型的权重。

[0078] 在框356，该系统确定是否基于附加训练示例继续训练语言选择模型。如果该系统在框356确定基于附加训练示例继续训练神经网络，则该系统执行框354的另一次迭代(例如，框354A、354B和354C)。如果该系统在框356确定不基于附加训练示例继续训练语言选择模型，则该系统继续进行至框358并且结束训练。在一些实施方式中，该系统可以基于缺少附加训练示例而确定不继续训练语言选择模型。附加地或可替选地，该系统可以基于训练已经被执行了至少阈值时间量，训练已经被执行了至少阈值量的时期，确定语言选择模型的当前训练版本满足一个或多个标准和/或其它(多种)因素来确定不继续训练语言选择模型。

[0079] 现在参考图4和6，描绘了根据本文所公开的实施方式的用于使用(例如，使用图3的方法300训练的)经训练的语言选择模型的方法。在图4和5的描述之前，提供使用经训练的语言选择模型的实施方式的简要概述。在推导时，识别候选语言的子集$S\in\{1,\cdots,N\}$，其中N是所支持的话音辨识语言的集合，并且其中子集S是该系统将从其中选择给定语言的子集。S在本文也被称作M。子集S可以利用诸如本文所描述的那些的技术针对给定口头话语被确定(例如，基于在与给定口头话语相关联的用户简档中被指定的子集)。给定语言的预测可以被表示为：$y^{*}=\arg\max_{k\in\mathbf{S}} f(\mathbf{x};\mathbf{w})=\arg\max_{k\in\mathbf{S}}\mathbf{z}_{k}$。进一步地，从用户所接收的口头话语的长度可能有所变化。该口头话语可以被截取为固定持续时间的分段，并且每一个分段的部分可能重叠并且作为输入被提供至经训练的语言选择模型。该经训练语言选择模型的最终输出或预测概率可以是分段的重叠部分的平均，并且在等式5中表示：

[0080] $$y^{*}=\arg\max_{k\in\mathbf{S}}\mathbf{E}_{t}[f(\mathbf{x}^{t};w)]=\arg\max_{k\in\mathbf{S}}\mathbf{E}_{t}[\mathbf{z}_{k}^{t}] \quad \text{（等式5）}$$

[0081] 其中是$x^{t}$是作为第t个滑动窗口的输入分段的输入，并且$z^{t}$是来自经训练的语言选择模型的相对应响应。通过使用该滑动窗口方法，经训练的语言选择模型可以为长的口头话语提供更加鲁棒的系统。进一步地，该滑动窗口方法适用于本文所描述的若干感兴趣的用例(例如，关于图4和5)。这些用例是非限制性的并且出于示例性的目的在本文被公开。

[0082] 图4是示出根据本文所公开的实施方式的用于使用经训练的语言选择模型来选择话音辨识语言的示例方法400的流程图。为了方便，该流程图的操作参考执行该操作的系统来描述。该系统可以包括各种计算机系统的各种组件，诸如图1中所描绘的一个或多个组

件。此外，虽然方法400的操作是以特定顺序被示出，但是这并非意在作为限制。一个或多个操作可以被重新排序、被省略或者被添加。

[0083] 在框452，该系统接收捕获来自用户的当前口头话语的音频数据。例如，该音频数据可以经由计算设备的麦克风被捕获，并且可以捕获来自用户的西班牙语语言的口头话语。

[0084] 在框454，该系统提取当前口头话语的一个或多个特征。继续以上示例，该系统可以提取诸如梅尔频率倒谱系数(MFCC)、对数梅尔滤波器组(log-mel-filterbank)特征和/或其它特征的特征。进一步地，在框454，该系统可选地可以选择所提取特征的子集，其中所提取特征的子集包括高度指示与当前口头话语相对应的语言的特征。

[0085] 在框456，该系统使用经训练的语言选择模型处理当前口头话语的一个或多个特征以生成N种不同语言中的每一种语言的预测概率。继续以上示例，该系统可以处理当前口头话语以生成当前口头话语对应于西班牙语语言的第一预测概率0.8，生成当前口头话语的特征对应于英语语言的第二预测概率0.1，以及生成其余N种不同语言中的每一种语言的相对应概率。

[0086] 在框458，该系统基于在框456所生成的当前预测概率选择该N种不同语言中的当前口头语言。继续以上示例，与基于第二预测语言0.1选择英语语言或者基于选择任何其它语言的概率而选择它们相反，该系统可以基于指示当前口头语言对应于西班牙语的第一预测概率0.8来选择西班牙语语言。如本文所描述的，在各种实施方式中，该系统在框458基于与当前口头话语相关联的M种候选语言的预测概率来选择当前口头语言，其中M是N的子集。例如，M种候选语言的指示可以被包括在与音频数据一起被传输的数据中。基于这样的指示，可以基于仅针对M种候选语言所生成的概率来选择M种候选语言中的一种候选语言。在那些实施方式中，可以选择M种候选语言中最高概率的语言——即使存在针对作为N种语言中的一种语言但是不是M种语言中的一种语言的另一种语言的更高的概率。

[0087] 在框460，该系统基于在框458所选择的当前口头语言来选择话音辨识模型来执行音频数据的话音至文本(STT)处理。继续以上示例，该系统选择与西班牙语语言相关联的话音辨识模型。

[0088] 在框462，该系统确定使用所选择的话音辨识模型来执行STT处理，并且不使用任何未选择的话音辨识模型来执行STT处理。因此，关于未选择的话音辨识模型，该系统继续进行至框472，其中处理针对其它话音辨识模型而技术。

[0089] 在框464，该系统使用所选择的话音辨识模型对音频数据执行STT处理以确定对应于当前口头话语的一个或多个单词。继续以上示例，该系统使用西班牙语语言的话音辨识模型执行STT处理以确定西班牙语语言的口头话语的一个或多个单词。

[0090] 在框466，该系统生成响应于一个或多个单词的内容。继续以上示例，该系统生成响应于西班牙语语言的口头话语的内容。响应于口头话语的内容可以包括自然语言响应、搜索结果、通过与第三方代理交互所确定的内容、使得安装在计算设备或远程计算设备上的一个或多个应用启动的内容等。该系统可以在处理一个或多个单词以确定该一个或多个单词的意图以及可选地该意图的参数时利用自然语言理解(NLU)引擎和/或其它引擎，并且可以基于该意图和参数而生成响应内容。

[0091] 在框468，该系统提供内容以由计算设备渲染。继续以上示例，该系统可以提供内

容以经由计算设备向用户进行可听的和/或视觉的呈现。在附加地或可替选的实施方式中，在框464所生成的一个或多个单词可以在框466的变型处被确定为对应于控制智能设备的请求。在那些实施方式中，框468的变体可以包括将一个或多个命令直接提供至智能设备或者提供至控制该智能设备的第三方服务器，其中命令使得智能设备与请求一致地被控制。

[0092] 在框470，该系统确定用户是否已经提供了附加话语。如果该系统在框470接收到附加话语，则该系统可以返回至框452。如果该系统在框470未接收到来自用户的附加话语，则该系统可以继续进行至框472并且该处理结束。

[0093] 图5是示出根据本文所公开的实施方式的用于使用经训练的语言选择模型来选择话音辨识语言的另一种示例方法500的流程图。为了方便，流程图的操作参考执行该操作的系统来描述。该系统可以包括各种计算机系统的各种组件，诸如图1中所描绘的一个或多个组件。此外，虽然方法500的操作是以特定顺序被示出，但是这并非意在作为限制。一个或多个操作可以被重新排序、被省略或者被添加。

[0094] 在框552，该系统接收捕获来自用户的当前口头话语的音频数据。例如，该音频数据可以经由计算设备的麦克风被捕获，并且可以捕获来自用户的西班牙语语言的口头话语。

[0095] 在框554，该系统提取音频数据的一个或多个特征。继续以上示例，该系统可以提取诸如梅尔频率倒谱系数(MFCC)、对数梅尔滤波器组特征和/或其它特征的特征。

[0096] 在框556，该系统选择对应于第一候选口头语言的第一话音辨识模型和对应于第二候选口头语言的至少第二话音辨识模型以执行音频数据的STT处理。该系统可以继续进行至框562A和562B以开始音频数据的STT处理的执行。继续以上示例，该系统可以选择与西班牙语语言相关联的第一话音辨识模型以用于音频数据的STT处理，并且选择与英语语言相关联的第二话音辨识模型以用于音频数据的STT和处理。进一步地，该系统可以使用西班牙语语言模型和英语语言模型来执行音频数据的STT处理。如本文所描述的，在各种实施方式中，该系统在框556基于第一和第二话音辨识模型针对与当前口头话语相关联的M种候选语言中的相应候选语言来选择第一和第二话音辨识模型，其中M是N的子集。例如，英语和西班牙语话音辨识模型可以基于那些是针对所接收音频数据的两种候选语言的指示而被选择并且用于STT处理。

[0097] 在框558，该系统使用经训练的语言选择模型处理该音频数据的一个或多个特征以生成N种不同语言中的每一种语言的预测概率。继续以上示例，该系统可以处理当前口头话语来以生成当前口头话语对应于西班牙语语言的第一预测概率0.4，生成当前口头话语的特征对应于英语语言的第二预测概率0.1，以及生成其余N种不同语言中的每一种语言的相对应概率。如本文所描述的，在各种实施方式中，框558与框562A和框562B的执行的至少一部分并行地执行。换句话说，针对M种候选语言中的每一种候选语言的STT处理可以在生成概率以使得能够选择口头话语的语言的同时被初始化。来自对应于所选择的语言的STT处理的输出然后可以被利用，并且可选地，针对其它语言的STT处理可以在这样的处理在已经选择了口头话语的语言时尚未完成的情况下被停止。通过执行这样的并行处理，生成对应于口头话语的文本时的延时可以减少，并且作为结果，可以以减少的延时基于口头话语采取响应性动作。进一步地，在针对其它(未确定的)语言的STT处理被停止的实施方式中，可以防止在这样的处理中所采用的不必要的资源消耗。

[0098] 在框560,该系统基于预测概率选择N种不同语言中的口头语言。继续以上示例,该系统可以基于第一预测概率0.4是M种候选语言的全部概率中的最高概率来选择西班牙语语言作为口头语言话语。

[0099] 在框562A,该系统使用对应于第一候选口头语言的第一话音辨识模型执行音频数据的STT处理以确定对应于当前口头话语的一个或多个单词。继续以上示例,该系统使用西班牙语语言模型执行音频数据的STT处理以确定对应于当前口头话语的西班牙语语言的一个或多个单词。

[0100] 在框562B,该系统使用对应于第二候选口头语言的至少第二话音辨识模型来执行音频数据的STT处理以确定对应于当前口头话语的一个或多个单词。继续以上示例,该系统使用英语语言模型来执行音频数据的STT处理以确定对应于当前口头话语的英语语言的一个或多个单词。

[0101] 如上文所描述的,在各种实施方式中,框562A和562B的STT处理可以与框558和560的处理并行地执行。例如,在使用语言选择模型执行处理的同时,可以针对M种候选语言(例如,继续示例中的西班牙语语言和英语语言)中的每一种候选语言初始化STT处理。在框560选择的口头语言被用来选择由框562A和562B中的仅一个所生成的输出(即使用对应于所选择的口头语言的话音辨识模型所生成的输出)。进一步地,如果利用未选择语言的STT处理尚未结束,则该系统可选地可以在框560选择口头语言之后停止这样的处理。

[0102] 继续之前的示例,该系统可以使用西班牙语语言模型和英语语言模型两者来执行音频数据的STT处理。在该系统执行此STT处理时,该系统可以使用语言选择模型来处理相对应的音频数据以生成N种不同语言中的每一种语言的概率。基于西班牙语语言的预测概率为0.4且英语语言的预测概率为0.2,该系统可以选择西班牙语语言并且使用用西班牙语辨识模型所生成的输出。该系统可选地可以在使用英语语言模型的STT处理在选择了西班牙语时尚未完成的情况下将该STT处理停止。在一些实施方式中,停止可选地可以仅在西班牙语语言的概率满足了阈值(例如,相对于英语语言概率的阈值)的情况下发生。例如,如果预测概率相对接近(例如,西班牙语语言的0.55和英语语言的0.45),则该系统可以完成使用西班牙语语言模型和英语语言模型二者的STT处理,并且使用来自STT处理的置信度量度和/或其它量度来确保所选择的西班牙语语言实际上是正确的语言。

[0103] 在框564,该系统确定使用对应于所选择语言的模型的STT处理是否完成。如果该系统在框564确定该STT处理未完成,则该系统继续框562A和/或562B的STT处理。如果该系统在框564确定该STT处理完成,则该系统继续进行至框566。在框566,该系统生成响应于使用所选择语言的STT处理所生成的一个或多个单词的内容。继续以上示例,该系统生成响应于西班牙语语言的口头话语的内容。

[0104] 在框568,该系统提供内容以由计算设备渲染。虽然关于利用两个模型(即,其中M=2)执行STT处理描述了方法500,但是注意到,在M=3时可以针对三个相对应模型中的每一个相对应模型并行地执行STT处理,在M=4时可以针对四个相对应模型中的每一个相对应模型并行地执行STT处理等。

[0105] 图6是根据示例计算机系统610的框图。计算机系统610通常包括至少一个处理器614,该至少一个处理器经由总线子系统612与多个外围设备进行通信。这些外围设备可以包括例如包括存储器625和文件存储子系统626的存储子系统724、用户接口输出设备620、

用户接口输入设备622和网络接口子系统616。输入和输出设备允许用户与计算机系统610的交互。网络接口子系统616提供到外部网络的接口并且耦合至其它计算机系统中的相对应接口设备。

[0106] 用户接口输入设备622可以包括键盘、诸如鼠标、轨迹球、触摸板或图形板的指向设备、扫描仪、整合到显示器中的触摸屏、诸如语音辨识系统、麦克风的音频输入设备，和/或其它类型的输入设备。通常，术语“输入设备”的使用意在包括用于向计算机系统610中或通信网络上输入信息的所有可能类型的设备和方式。

[0107] 用户接口输出设备620可以包括显示子系统、打印机、传真机，或者诸如音频输出设备的非视觉显示器。显示子系统可以包括阴极射线管(CRT)、诸如液晶显示器(LCD)的平板设备、投影设备，或者用于创建可视图像的一些其它机制。显示子系统还可以诸如经由音频输出设备提供非视觉显示器。通常，术语“输出设备”的使用旨在包括用于从计算机系统610向用户或者向另一个机器或计算机系统输出信息的所有可能类型的设备和方式。

[0108] 存储子系统624存储提供本文所描述的一些或全部模块的功能的编程和数据构造。例如，存储子系统624可以包括用于执行方法300、400和500的所选择的方面和/或用于实现服务器设备102、客户端计算设备118、便携式计算设备132和/或本文所讨论的任何其它设备或操作的逻辑。

[0109] 这些软件模块通常由处理器614单独地或者结合其它处理器来执行。存储子系统624中使用的存储器625可以包括多个存储器，该多个存储器包括用于程序执行期间的指令和数据的存储的主随机访问存储器(RAM)630以及其中存储固定指令的只读存储器(ROM)632。文件存储子系统626可以为程序和数据文件提供永久存储，并且可以包括硬盘驱动器、软盘驱动器、连同相关联的可移除介质一起、CD-ROM驱动器、光驱或可移除介质卡盒。实现某些实施方式的功能的模块可以由文件存储子系统626存储在存储子系统624或者能够由处理器614所访问的其它机器中。

[0110] 总线子系统612提供用于使得计算机系统610的各个组件和子系统如所期望的互相通信的机制。虽然总线子系统612被示意性地示为单总线，但是总线子系统的可替选实施方式可以使用多个总线。

[0111] 计算机系统610可以是变化的类型，包括工作站、服务器、计算集群、刀片服务器、服务器场，或者任何其它的数据处理系统或计算设备。由于计算机和网络的不断变换的本质，图6中所描绘的计算机系统610的描述仅意在作为出于说明一些实施方式的目的的具体的示例。具有与图6中所描绘的计算机系统相比更多或更少组件的计算机系统610的许多其它配置是可能的。

[0112] 在本文所描述的系统收集有关用户(或者如本文经常所提到的“参与者”)的个人信息或者可以利用个人信息的情况下，用户可以被提供控制程序或特征是否收集用户信息(例如，有关用户的社交网络、社交动作或活动、职业、用户的偏好或用户的当前地理位置的信息)，或者控制是否和/或如何从内容服务器接收可能与用户更为相关的内容的机会。而且，某些数据可以先于其被存储或使用而以一种或多种方式被处理，使得个人可识别信息被移除。例如，用户的身份可以被处理使得不能确定用户的个人可识别信息，或者可以在获得位置信息的情况下对用户的地理位置进行一般化处理(诸如一般化为城市、邮政编码或州级)，使得用户的特定位置不能被确定。因此，用户可以控制如何收集和/或使用有关该用

户的信息。

[0113] 虽然本文已经描述和示出了若干实施方式，但是可以利用用于执行本文所描述的功能和/或获得本文所描述的结果和/或一个或多个优势的各种其它手段和/或结构，并且这样的变体和/或修改均被认为在本文所描述的实施方式的范围之内。更一般地，本文所描述的所有参数、尺寸、材料和配置都意在是示例性的，并且实际的参数、尺寸、材料和/或配置将取决于使用本教导的一个或多个具体应用。本领域技术人员仅使用常规实验就将认识到或者能够确认本文所描述的具体实施方式的许多等同物。因此，要理解的是，前述实施方式仅通过示例给出，并且在所附权利要求及其等同物的范围内，可以以与具体描述和要求保护的方式不同的方式实施。本公开的实施方式涉及本文所描述的每一个个体特征、系统、物品、材料、装备和/或方法。此外，如果这样的特征、系统、物品、材料、装备和/或方法不互相矛盾，则两个或更多个这样的特征、系统、物品、材料、装备和/或方法的任意组合被包括于本公开的范围之内。

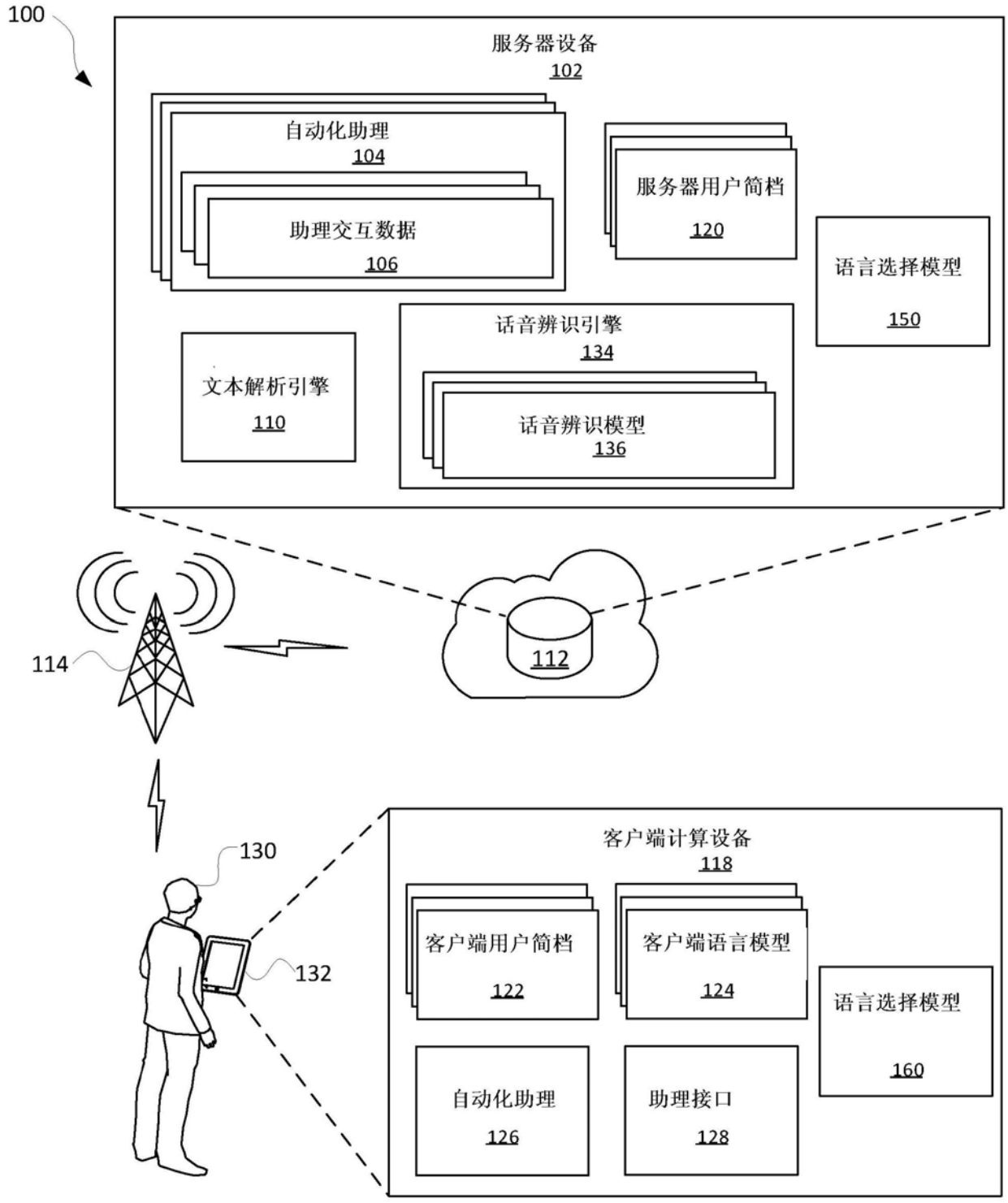

100
服务器设备
102
自动化助理
104
助理交互数据
106
服务器用户简档
120
语言选择模型
150
话音辨识引擎
134
话音辨识模型
136
文本解析引擎
110
114
112
130
132
客户端计算设备
118
客户端用户简档
122
客户端语言模型
124
语言选择模型
160
自动化助理
126
助理接口
128

图1

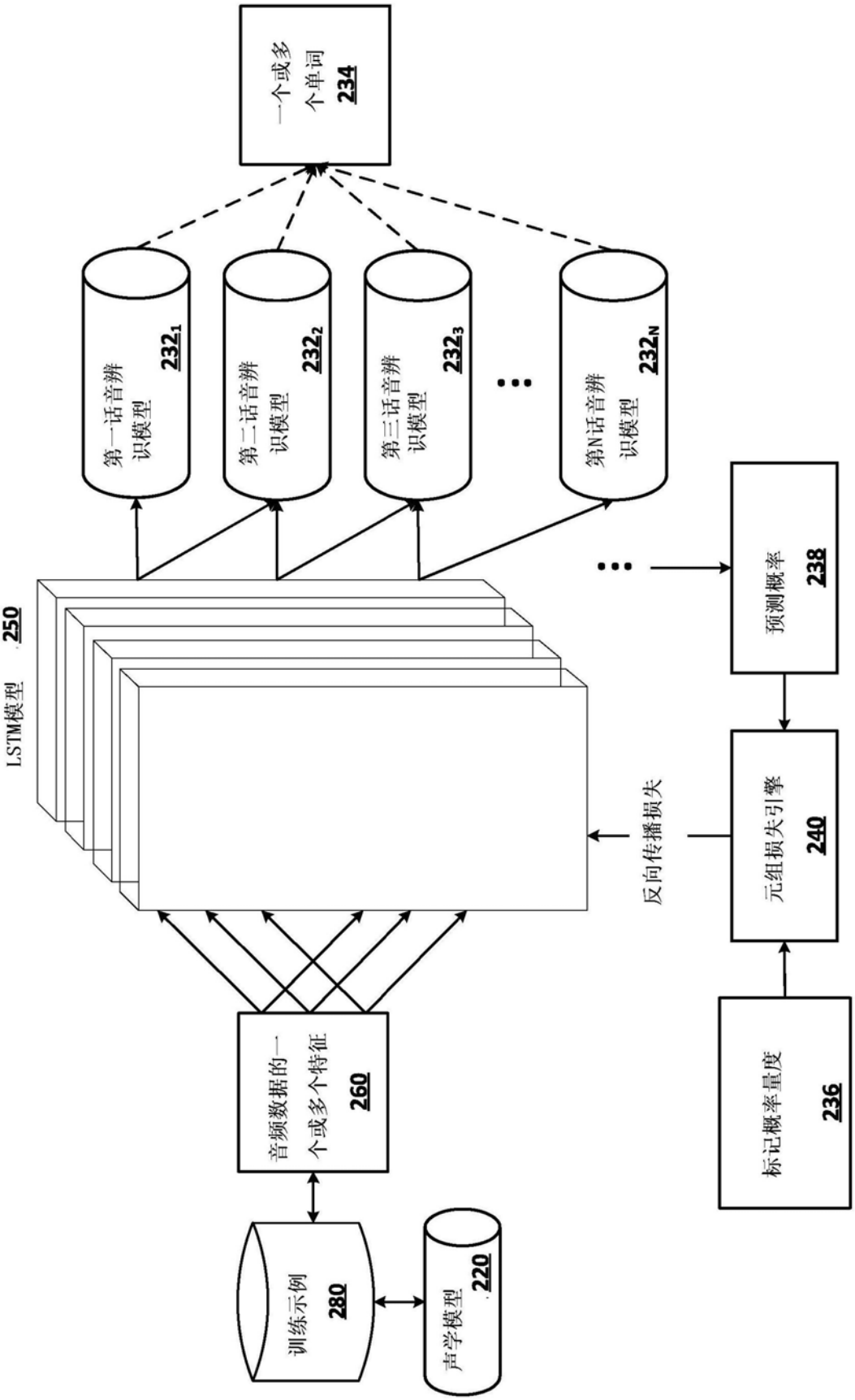

一个或多个单词 234
第一话音辨识模型 232₁
第二话音辨识模型 232₂
第三话音辨识模型 232₃
第N话音辨识模型 232N
LSTM模型 250
预测概率 238
元组损失引擎 240
反向传播损失
标记概率量度 236
音频数据的一个或多个特征 260
训练示例 280
声学模型 220

图2

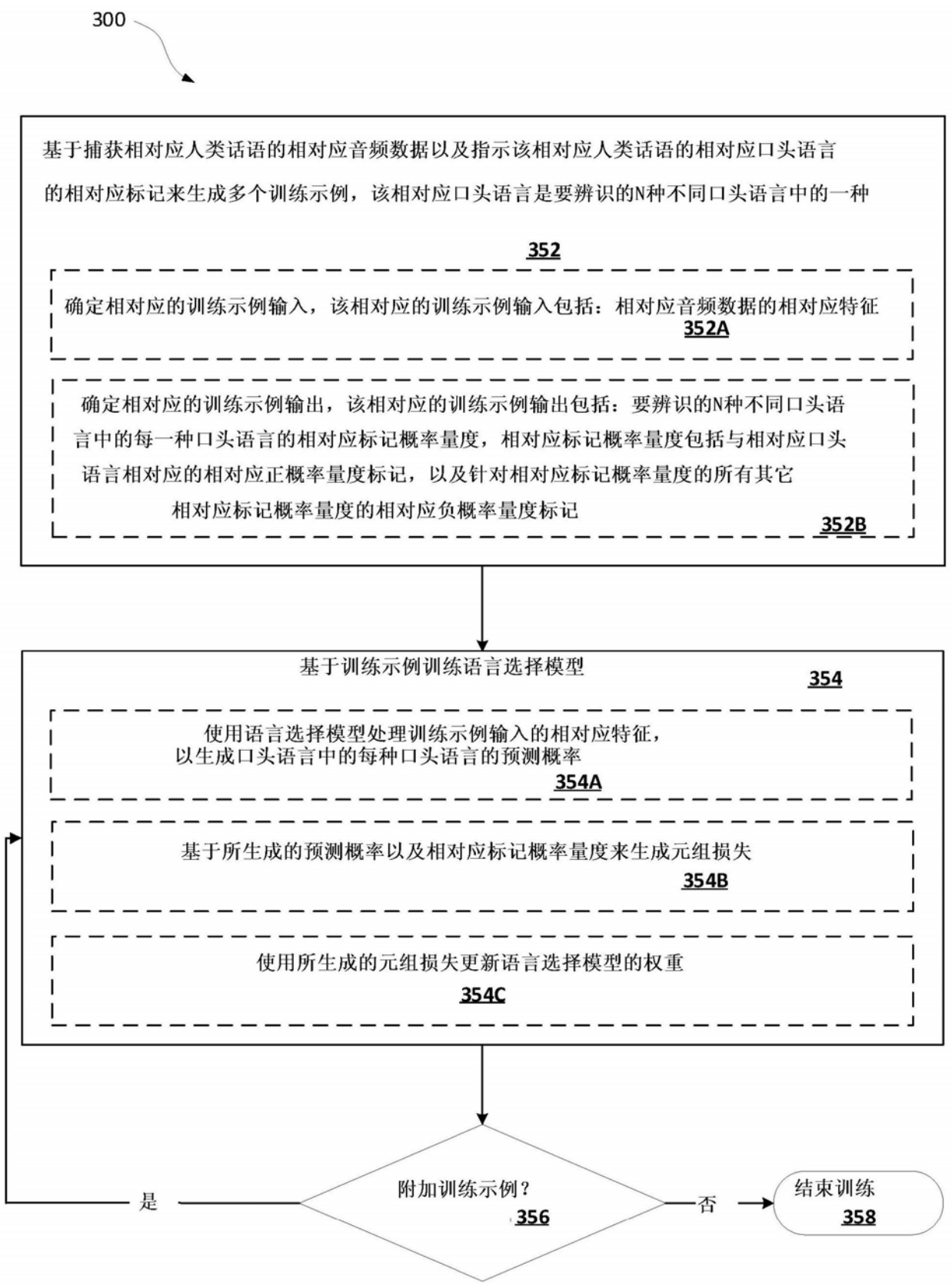

300
基于捕获相对应人类话语的相对应音频数据以及指示该相对应人类话语的相对应口头语言的相对应标记来生成多个训练示例，该相对应口头语言是要辨识的N种不同口头语言中的一种
352
确定相对应的训练示例输入，该相对应的训练示例输入包括：相对应音频数据的相对应特征
352A
确定相对应的训练示例输出，该相对应的训练示例输出包括：要辨识的N种不同口头语言中的每一种口头语言的相对应标记概率量度，相对应标记概率量度包括与相对应口头语言相对应的相对应正概率量度标记，以及针对相对应标记概率量度的所有其它相对应标记概率量度的相对应负概率量度标记
352B
基于训练示例训练语言选择模型
354
使用语言选择模型处理训练示例输入的相对应特征，以生成口头语言中的每种口头语言的预测概率
354A
基于所生成的预测概率以及相对应标记概率量度来生成元组损失
354B
使用所生成的元组损失更新语言选择模型的权重
354C
附加训练示例？
356
是
否
结束训练
358

图3

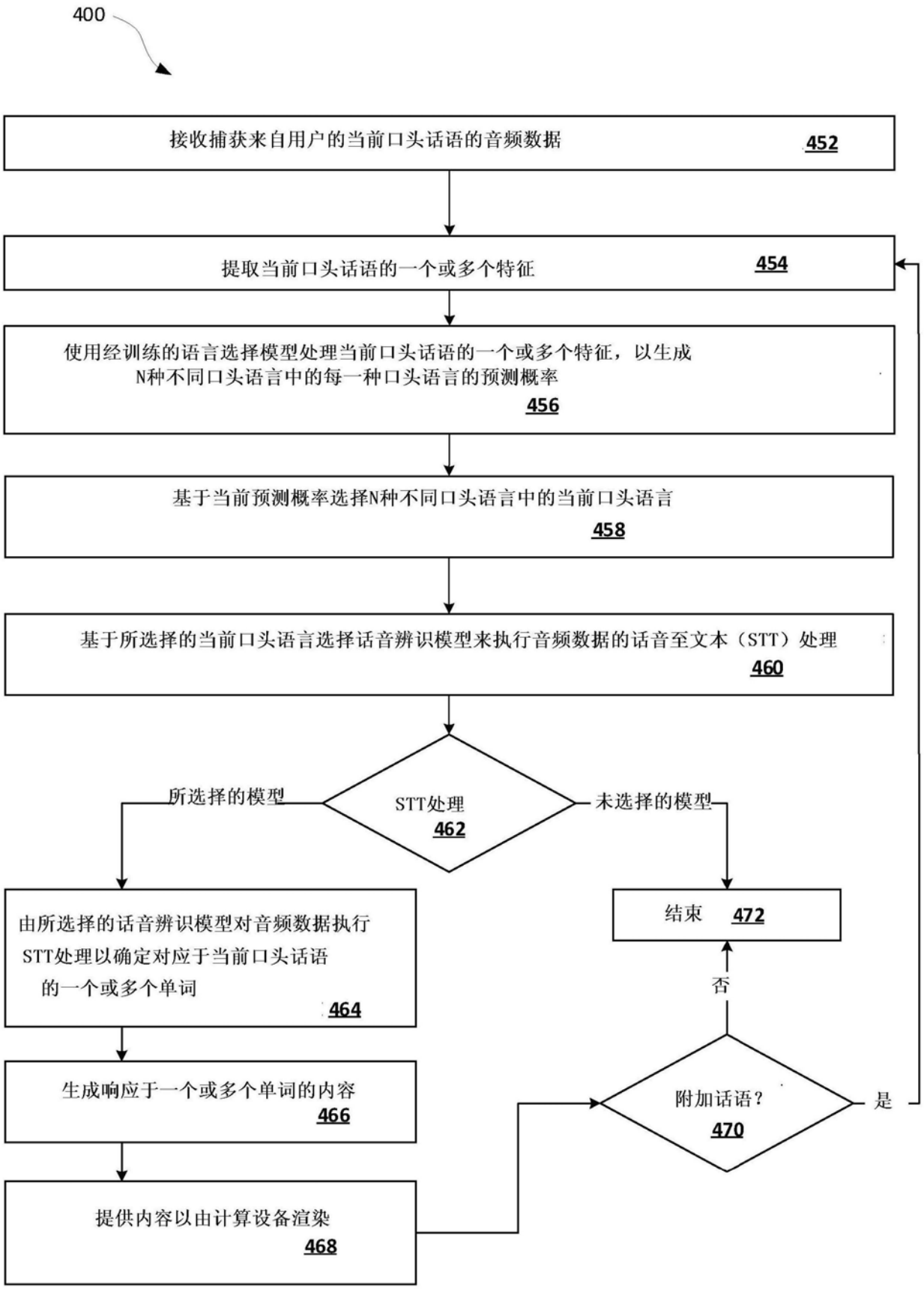

400
接收捕获来自用户的当前口头话语的音频数据
452
提取当前口头话语的一个或多个特征
454
使用经训练的语言选择模型处理当前口头话语的一个或多个特征，以生成N种不同口头语言中的每一种口头语言的预测概率
456
基于当前预测概率选择N种不同口头语言中的当前口头语言
458
基于所选择的当前口头语言选择话音辨识模型来执行音频数据的话音至文本（STT）处理
460
STT处理
462
所选择的模型
未选择的模型
由所选择的话音辨识模型对音频数据执行STT处理以确定对应于当前口头话语的一个或多个单词
464
生成响应于一个或多个单词的内容
466
提供内容以由计算设备渲染
468
结束
472
否
附加话语？
470
是

图4

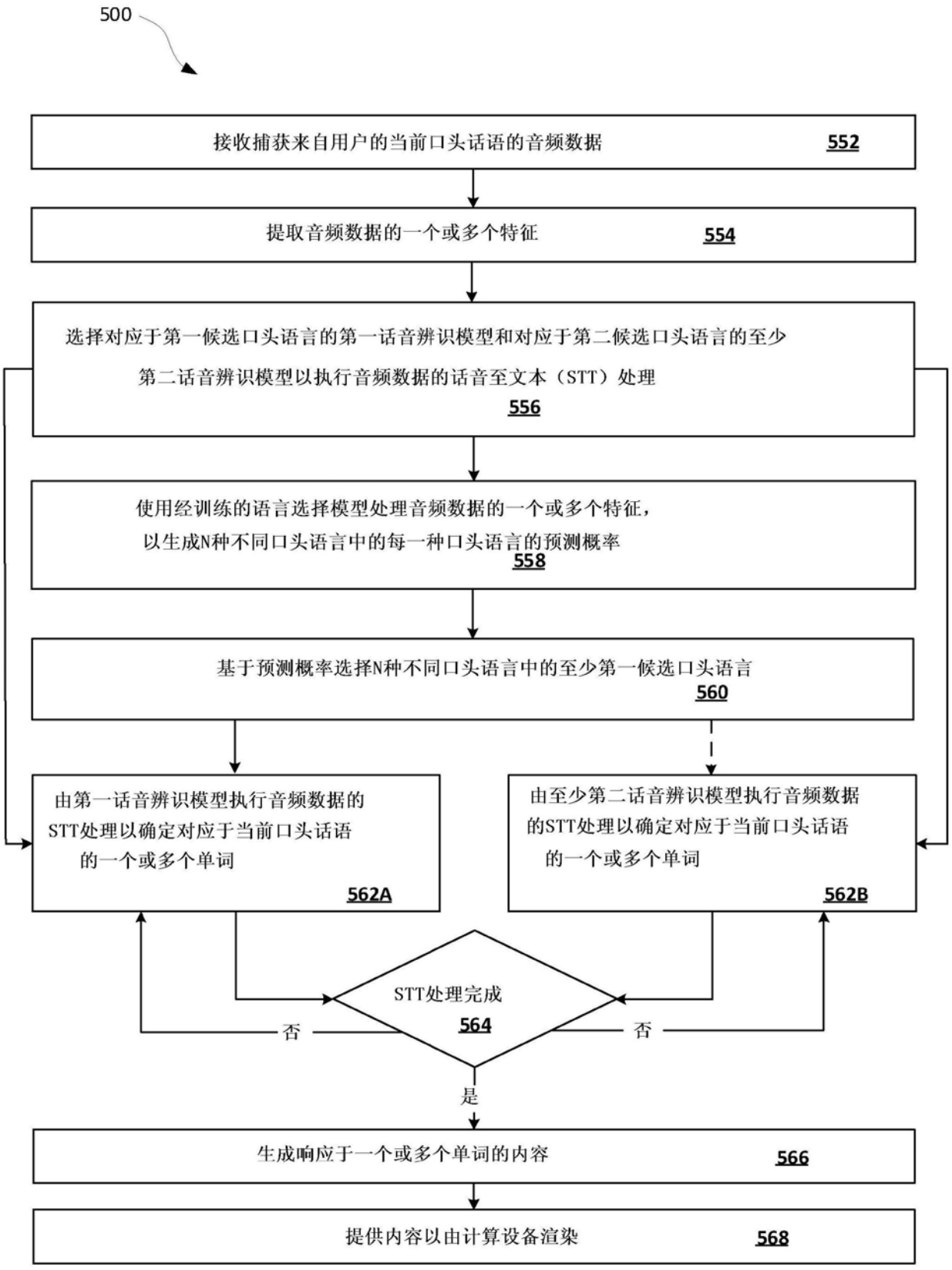

500
接收捕获来自用户的当前口头话语的音频数据 552
提取音频数据的一个或多个特征 554
选择对应于第一候选口头语言的第一话音辨识模型和对应于第二候选口头语言的至少第二话音辨识模型以执行音频数据的话音至文本（STT）处理 556
使用经训练的语言选择模型处理音频数据的一个或多个特征，以生成N种不同口头语言中的每一种口头语言的预测概率 558
基于预测概率选择N种不同口头语言中的至少第一候选口头语言 560
由第一话音辨识模型执行音频数据的STT处理以确定对应于当前口头话语的一个或多个单词 562A
由至少第二话音辨识模型执行音频数据的STT处理以确定对应于当前口头话语的一个或多个单词 562B
STT处理完成 564
否
否
是
生成响应于一个或多个单词的内容 566
提供内容以由计算设备渲染 568

图5

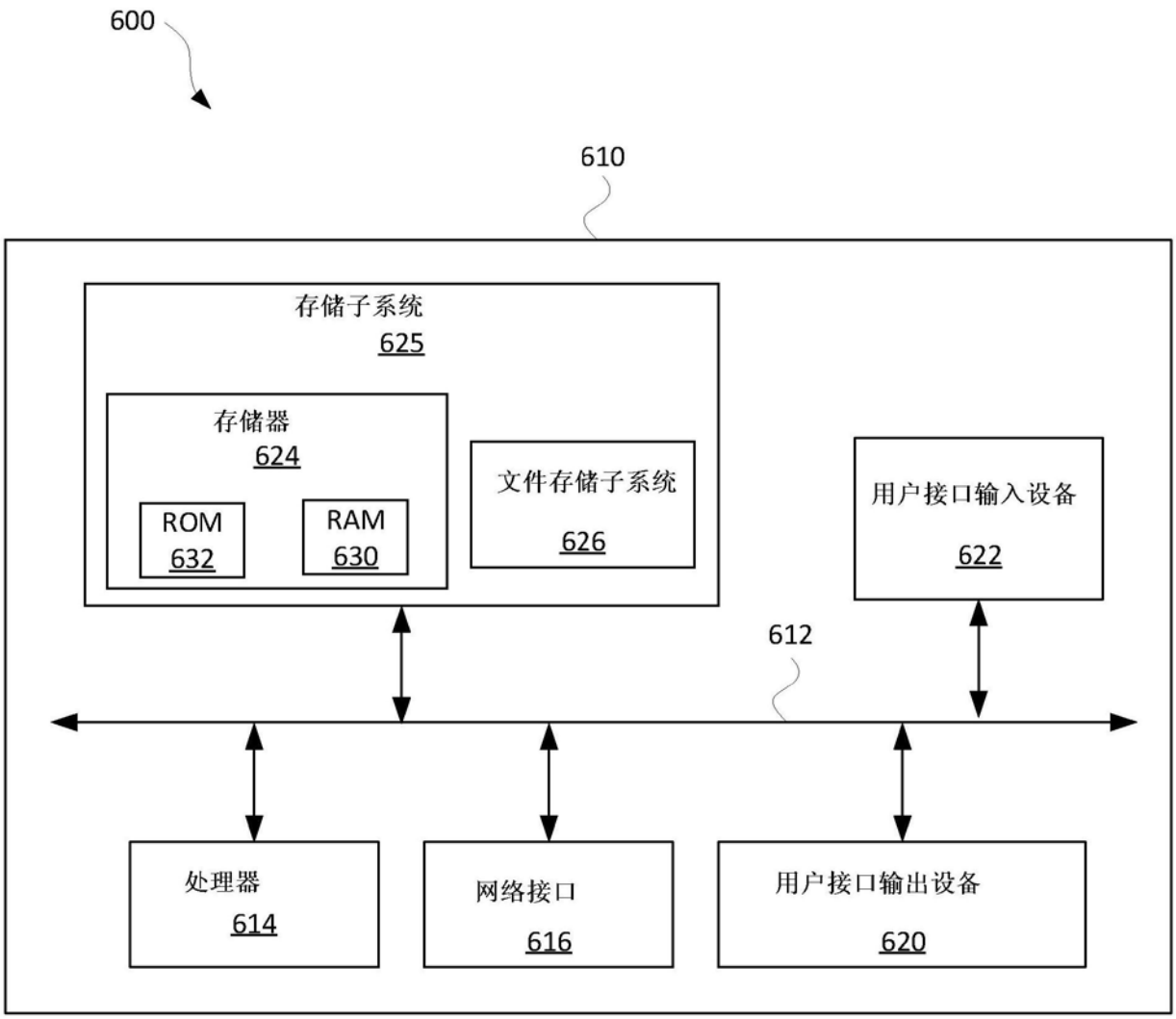

600
610
存储子系统
625
存储器
624
ROM
632
RAM
630
文件存储子系统
626
用户接口输入设备
622
612
处理器
614
网络接口
616
用户接口输出设备
620

图6