(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 104115221 B
(45)授权公告日 2017.09.01

(21)申请号 201380009453.4

(22)申请日 2013.02.01

(65)同一申请的已公布的文献号
申请公布号 CN 104115221 A

(43)申请公布日 2014.10.22

(30)优先权数据
13/399,496 2012.02.17 US

(85)PCT国际申请进入国家阶段日
2014.08.14

(86)PCT国际申请的申请数据
PCT/US2013/024245 2013.02.01

(87)PCT国际申请的公布数据
WO2013/122750 EN 2013.08.22

(73)专利权人 微软技术许可有限责任公司
地址 美国华盛顿州

(72)发明人 Y·钱 B·B·朱 F·K-P·宋

(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 13/033(2013.01)

(56)对比文件
US 2010/0312562 A1,2010.12.09,
US 2012/0004914 A1,2012.01.05,
CN 1758330 A,2006.04.12,
CN 1584979 A,2005.02.23,
郭峰.《汉语语音验证码技术及应用》.《中国优秀硕士学位论文全文数据库 信息科技辑》.2011,(第04期),
袁俏.《一种智能变换语音技术的研究》.《中国优秀硕士学位论文全文数据库》.2011,(第S1期),
Greg Kochanski et al..《A REVERSE TURING TEST USING SPEECH》.《ICSLP 2002:7th International Conference On Spoken Language Processing》.2002,

审查员 李海龙

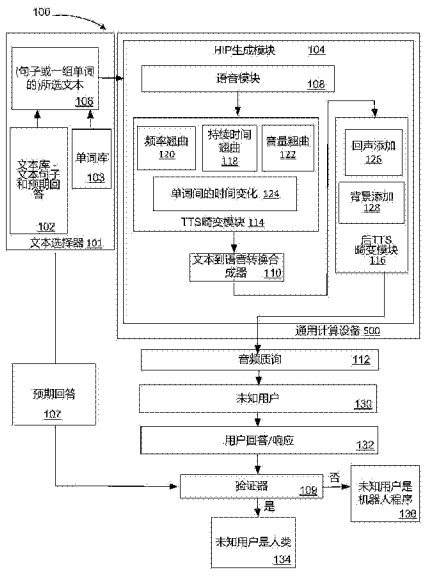
权利要求书1页 说明书11页 附图5页

(54)发明名称

基于文本到语音转换以及语义的音频人类交互证明

(57)摘要

本文所述的文本到语音转换音频HIP技术在一些实施例中使用经由文本到语音转换引擎生成的不同的相关或不相关的单词或句子作为音频HIP质询。该技术可在说出要被用作HIP质询串的句子的文本到语音转换合成器中应用不同的效应。这些不同的效应可包括例如谱频率翘曲;元音持续时间翘曲;背景添加;回声添加;以及改变单词间的持续时间等等。在某些实施例中,该技术改变该组参数来防止自动语音识别工具使用先前使用的音频HIP质询来习得可然后被用来识别由该技术生成的将来的音频HIP质询的模型。此外,在一些实施例中,该技术在HIP质询中引入语义理解的要求。



CN 104115221 B

1.一种用于提供自动人类交互证明的计算机实现的方法,包括:

从多个文本句子或多个单词中选择文本串;以及

应用文本到语音转换引擎来生成所选文本的话音作为音频质询,包括应用以下各项中的一者或多者来生成所选文本的话音:谱频率翘曲、元音持续时间翘曲、音量翘曲、背景添加、回声添加、以及单词间持续时间,所述音频质询要求对所选文本的语义知识来对所述音频质询作出回答,用于标识未知用户是人类还是机器人程序,这些参数集合被改变以禁用使用先前使用的音频质询以使其不能被用于训练以识别所生成的音频质询。

2.如权利要求1所述的计算机实现的方法,其特征在于,进一步包括:

从所述未知用户接收对所述音频质询的回答;

验证来自所述未知用户的对所述音频质询的所述回答,以确定所述未知用户是人类还是机器人程序。

3.如权利要求2所述的计算机实现的方法,其特征在于,接收到的回答被所述未知用户说出,并且其中话音识别被应用来识别该回答并将该回答与正确的答案相比较。

4.如权利要求1所述的计算机实现的方法,其特征在于,需要对所述音频质询的语义理解来达到正确的答案。

5.一种用于生成用于自动人类交互证明的基于音频的质询的系统,包括:

通用计算设备;

包括可由所述通用计算设备执行的程序模块的计算机程序,其中所述计算机程序的所述程序模块指示所述计算设备来:

选择要被用作质询串的文本串,对于所述质询串,对所述质询作出回答需要对所述文本串的语义理解;

用一个或多个畸变来对所选文本的语音模型的参数进行畸变,使得所选文本句子在被文本到语音转换引擎读出时将被扭曲,包括应用以下各项中的一者或多者:谱频率翘曲、元音持续时间翘曲、音量翘曲、背景添加、回声添加、以及单词间持续时间,这些参数集合被改变以禁用使用先前使用的音频质询以使其不能被用于训练以识别所生成的音频质询;

使用被畸变的参数以及所述语音模型,使用文本到语音转换合成器来读出所选文本句子作为对未知计算机用户的音频质询;以及

自动确定来自所述未知计算机用户的回答是否与预期回答相匹配。

6.如权利要求5所述的系统,其特征在于,所述未知计算机用户必须具有对所选文本串的语义理解,以便提供预期回答。

基于文本到语音转换以及语义的音频人类交互证明

背景技术

[0001] 人类交互证明(HIP)，也称为CAPTCHA(区分计算机和人类的完全自动化公共图灵测试)，它将人类用户与自动编程(即机器人程序(bot))相区分。

[0002] 大多数HIP方案的目标是防止计算机自动访问，而允许人类访问。通常，该目标是通过提供一种用于生成和分级大多数人能够容易通过而大多数计算机程序不能通过的测试来解决的。

发明内容

[0003] 提供本发明内容是为了以简化的形式介绍将在以下详细描述中进一步描述的一些概念。本发明内容并不旨在标识所要求保护主题的关键特征或必要特征，也不旨在用于限制所要求保护主题的范围。

[0004] 本文描述的文本到语音转换音频人类交互证明(HIP)技术提供了一种音频HIP，它在创建音频质询时采用文本到语音转换技术以及语义，以确定未知计算机用户是人类还是机器人程序。为了使得说出的句子不能被通用或定制的自动语音识别(ASR)系统(通常由机器人程序使用来尝试自动地解密音频HIP)识别，该技术防止ASR系统的语音识别机制识别由该技术生成的HIP样本或从中学习。该技术通过使得说出的HIP句子非常不同于在训练ASR系统的模型时使用的音频数据，以及通过改变说出的HIP单词或句子中的特征来实现这一点。ASR系统通常基于统计模型。HIP句子离ASR模型的训练数据分布越远，ASR系统越难以识别HIP的单词或句子。

[0005] 本文所述的文本到语音转换音频HIP技术可在通过文本到语音转换(TTS)引擎生成说出的HIP句子时应用谱频率翘曲、元音持续时间翘曲、音量翘曲、背景添加、回声添加、以及单词间持续时间。所得到的说出的句子的节奏、音高、以及话音因此与用于训练ASR系统的常用数据非常不同。此外，该技术采 用大集合的文本到语音转换参数，来允许该技术频繁地或经常地改变各种效应以禁止使用之前使用的音频HIP质询，使其不能被用于训练ASR系统的模型以识别该技术所生成的HIP质询。

[0006] 本文所述的文本到语音转换音频HIP技术的一个实施例可添加附加机制来将人类用户与机器人程序相区分：音频HIP质询的句子必须被理解以便通过该质询。例如，该句子可以是一问题或指令，需要对作为音频质询提出的该句子的语义理解以便正确地回答该质询。以此方式，即使先前描述的机制失败，即ASR系统能够识别被用作音频质询的句子中的全部单词，在不理解该句子的情况下机器人程序可能仍然不能通过该测试。对句子的语义理解仍然被认为是一种挑战性的人工智能问题。

附图说明

[0007] 参考以下描述、所附权利要求书以及附图，将更好地理解本公开的具体特征、方面和优点，附图中：

[0008] 图1是用于实践本文所描述的文本到语音转换音频HIP技术的一个示例性实施例

的示例性架构。

[0009] 图2描绘了用于采用文本到语音转换音频HIP技术的一个实施例的示例性过程的流程图。

[0010] 图3描绘了用于采用文本到语音转换音频HIP技术的一个实施例的示例性过程的另一流程图。

[0011] 图4描绘了用于采用文本到语音转换音频HIP技术的一个实施例的示例性过程的又一流程图。

[0012] 图5是可用于实践文本到语音转换音频HIP技术的示例性计算环境的示意图。

具体实施方式

[0013] 在以下对文本到语音转换音频HIP技术的描述中，对附图作出参考，附图形成了该描述的一部分，且通过可实践此处所描述的文本到语音转换音频HIP技术的说明性示例示出。要理解，可以利用其他实施例，并且可以作出结构上 的改变而不背离所要求保护的主题的范围。

[0014] 1.0文本到语音转换音频HIP技术

[0015] 以下各节提供了对人类交互证明(HIP)的介绍、此处描述的文本到语音转换音频HIP技术的概览、以及用于实践该技术的示例性架构和示例性过程。还提供了该技术的各种实施例的细节。

[0016] 1.1人类交互证明(HIP)的介绍

[0017] HIP，也称为CAPTCHA(区分计算机和人类的完全自动化公共图灵测试)，它将人类用户与自动编程(即机器人程序(bot))相区分。大多数HIP方案的目的是防止计算机自动访问，而允许人类访问。通常，该目标是通过提供一种用于生成和分级大多数人能够容易通过而大多数计算机程序不能通过的测试来解决的。

[0018] 目前有许多HIP方案可用。例如，一种常规视觉方案通过从字典随机选择字符或单词、然后呈现包含所述字符或单词的畸变图像来操作。该方案然后向其用户提供一测试，该测试由该畸变图像以及要键入出现在该图像中的一些字符或单词的请求所组成。通过定制所应用的变形的类型，创建一图像，其中大多数人能从畸变图像中读取所需数量的字符或单词，而当前的计算机程序通常不能。

[0019] 在另一音频HIP示例中，各个字符由人说出。与伴随的视觉HIP相同的那些说出的字符被扭曲并且被拼接在一起，其中在字母间有不同的时间持续期。还添加了背景噪声。用户被要求键入说出的字母。

[0020] 在又一音频HIP中，各个单词被说出、变形且加有背景噪声。用户被要求键入说出的单词。所键入的单词中存在一些误差是可容忍的。

[0021] 1.2技术概览

[0022] 本文所述的文本到语音转换音频HIP技术在一些实施例中使用经由文本到语音转换引擎生成的不同的(最好是不重复的)句子或单词作为音频HIP质询。该技术可在说出要被用作HIP的句子或单词的文本到语音转换合成器中施加不 同的效应。这些不同的效应可包括例如谱频率翘曲；元音持续时间翘曲；音量翘曲；背景添加；回声添加；单词间的持续时间的变化等等。在某些实施例中，该技术改变参数集合来随时间生成音频HIP质询并用于生

成不同的质询，以便阻止ASR习得可被用来识别该技术所生成的音频HP质询的模型。此外，在一些实施例中，为了解答HIP质询，该技术引入语义理解的要求。

[0023] 1.3示例性架构

[0024] 图1示出用于实践文本到语音转换音频HIP技术的一个实施例的示例性架构100。如图1所示，该示例性架构100包括文本选择器模块101，文本选择器模块可包含文本库102(例如文本句子和预期回答)或单词库103。模块101选择文本106并将其提供给HIP生成模块104，以及选择预期回答107并将其提供给验证器109。在一个实施例中，文本选择器模块101可以随机方式或特定方式来选择特定项(如最好是文本句子及其预期回答)。文本库102中的句子可从由特定源提供的文档或文章中选择、从因特网爬寻、或基于特定规则或模式从一模块(图1中未示出)生成。在某些实施例中，预期回答是与句子一起产生的。预期回答可以是句子本身，或是人类响应于该句子而将提供的回答。前者通常被用在句子是从文档或文章自动产生的时候。后者通常被用在句子是由程序模块产生的时候。

[0025] 在一个实施例中，文本选择器模块110可包含可从中以随机方式或特定方式选择一组相关或不相关的单词的单词库。所选单词被用作被发送给HIP生成模块104的所选文本106，与所选文本106相同顺序排列的单词还被用作被发送给验证器109的预期回答107。

[0026] 该架构包括与文本选择器101驻留在同一或不同通用计算设备500上的音频HIP生成模块104。将参考图5来更详细地讨论通用计算设备500。HIP生成模块104包含TTS引擎、TTS畸变模块114以及后TTS畸变模块116。一种常规的TTS引擎由两部分组成：语音模型108以及文本到语音转换合成器110。该TTS引擎使用语音模型108来处理所选文本106。在两个TTS引擎组成部分(语音模型108和TTS合成器110)之间，TTS畸变模块114调整由语音模型108建模的参数来应用一个或多个畸变，使得所选文本106将在被文本到语音 转换合成器110读出时被扭曲。TTS输出可进一步由后TTS畸变模块116来处理，以应用一个或多个附加畸变，例如向TTS生成的说出的文本添加回声或背景。所产生的声音被用作音频HIP/CATCHA 112。对于生成音频质询串的每个实例而言，这些畸变或者定义了TTS合成器110或后TTS中的一个或多个畸变的参数可随机地或以特定模式改变。

[0027] HIP生成模块104确定被用于使用语音模型108来对所选文本进行建模的畸变参数。在一个实施例中，该语音模型108是被用于对频谱(声道)、基频(声源)以及语音的持续时间(韵律)进行建模的隐形马尔可夫模型(HMM)。HIP生成模块104内部的TTS畸变模块114可包括频率翘曲模块120，频率翘曲模块120在所选文本106被文本到语音转换合成器110读出时翘曲所选文本的频率参数。TTS畸变模块114还可包括用于改变可发音的声音的持续时间的模块118。例如，该模块118可执行元音持续时间翘曲，元音持续时间翘曲在所选句子106被文本到语音转换合成器110读出时改变所选句子106的元音的发音时间。此外，TTS畸变模块114可包括在文本到语音转换合成器110为所选文本106生成话音时用于改变声音音量的模块122和/或改变单词间的持续时间的模块124。

[0028] 在文本到语音转换合成器110生成了所选文本的话音之后，可利用后TTS畸变模块116应用一个或多个附加畸变。后TTS畸变模块116可包括添加回声效应的回声添加模块126和/或向从文本到语音转换合成器110生成的所选文本106的音频剪辑添加背景声音的背景添加模块128。

[0029] 背景添加模块128可添加不同的背景声音。在一个实施例中，可添加音乐作为背景

声音。在另一实施例中，可向所选文本106的来自文本到语音转换合成器110的话音(称为前景话音)添加另一话音(下文称为背景话音)。可向背景声音应用畸变和其他修改，以对相同的或不同的音频HIP质询产生背景声音上的附加变化。

[0030] 当话音被添加时，背景话音可以与前景话音为相同的语言。它也可以是与前景话音的语言不同的语言。例如，当前景话音是英语时，背景话音可以是中文或西班牙语。背景话音可以与前景话音类似的方式用TTS合成器110生成。在背景话音的生成期间，可应用上面关于前景话音所提及的诸如频率翘曲等之 类的不同畸变。背景话音的文本可以是从文本库选择的句子或从字典随机选择的单词。有了添加的背景话音，人类能够容易地区分两种语言并标识和识别前景语言，而诸如ASR引擎之类的机器不能将前景话音与背景话音相区分，从而不能识别前景话音的说出的文本。

[0031] 从HIP生成模块生成的音频质询被发送给一未知用户130，该用户能用诸如使用键盘、鼠标、或触摸屏之类的各种方法来输入回答。在一个实施例中，该未知用户130可说出一回答，而话音识别技术可被用于识别该回答并将其转换成文本。接收到的文本回答132然后被发送给验证器109，验证器109将接收到的回答与音频质询的预期回答相比较。如果确定来自于该未知用户的回答132匹配预期回答107，则验证器109将该未知用户130标识为人类134。否则，该未知用户被标识为机器人程序136。在一个实施例中，该未知用户130识别音频质询112来提供正确的回答以便通过测试。在另一实施例中，为了提供正确的回答以便通过测试，该未知用户130必须对该音频质询具有语义理解。

[0032] 许多技术可被用在验证器109中来确定接收到的回答132是否匹配预期回答。在一个实施例中，验证器仅在这两个回答精确匹配时才确定他们彼此匹配。在该情况下，不容忍任何误差。在另一实施例中，如果两个回答之间的误差低于容限误差，则验证器确定这两个回答彼此匹配。在一个实施例中，两个回答之间的误差是使用编辑距离或其变体来计算的。

[0033] 验证器109可在将一回答与另一回答相比较之前处理该回答。例如，验证器109可对一回答的文本进行规范化(例如用其标准表达来替换一单词或一串文本)以及去除无关紧要的单词。验证器109也可将一文本回答转换成音素串，并对各音素串进行比较来确定两个回答是否彼此匹配。可使用许多技术来将文本转换成音素。在一个实施例中，TTS中的语音模型被用于将文本转换成音素。

[0034] 1.4用于实践该技术的示例性过程

[0035] 一般来说，图2示出用于实践本文所述的文本到语音转换音频HIP技术的一个实施例的一般示例性过程。如框202所示，从多个文本句子或多个单词选择文本句子或一组相关或不相关的单词。如框204所示，应用文本到语音转换引擎来生成所选文本的话音作为音频质询，用于标识未知用户是人类还是机器 人程序，其中在所选文本的话音的生成期间或之后应用了一个或多个畸变。

[0036] 图3示出用于实践文本到语音转换音频HIP技术的另一实施例的更详细的示例性过程300。一般来说，文本到语音转换音频HIP技术的该实施例通过首先查找或定义离散文本句子或单词的库来操作，如框302所示。例如，可从各种合适的文本源来选择文本句子。在一个实施例中，文本库中的文本是从人类容易理解的文章或文档(诸如面向普通人类读者的报纸和杂志)中自动提取的不重复的句子。在一个实施例中，提取的句子的长度具有预设范围。如果某一句子太短，则它可与下一句子组合或简单地被丢弃。太长的句子可被切成两

个或更多个更小的块以适合所要求的长度。文章或文档可从内部源提供，或者从因特网爬寻。在另一实施例中，该音频HIP技术构建或定义了单词库。可通过去除无关紧要的单词或令人混淆的单词来从一字典中构建该库。人类容易被其拼写或被其声音混淆的那些单词可从库中被移除。给定该文本句子库或单词库，如框304所示，该技术从文本句子库中自动选择一文本句子或从单词库中自动选择一组相关或不相关的单词，用于创建用来确定未知用户是人类还是机器人程序的音频质询。如框306所示，如果回答与被检索出的文本句子一起被存储在库中，则该技术还可从文本句子库检索预期回答，或者从所选的文本句子或该组相关或不相关的单词中生成预期回答。在一个实施例中，所生成的预期回答与所选文本串相同。在另一实施例中，所生成的预期回答是在文本规范化被应用于所选文本之后的结果。文本规范化接收一文本输入并产生一文本输出，该文本输出将该输入文本转换成标准格式。例如，在文本规范化期间，诸如“a”、“an”之类的无关紧要的单词可被移除，“I' m”可被“I am”替换。(如后面将讨论的，预期回答被发送给验证器316以与来自未知用户的回答314相比较，以确定该未知用户318是人类还是机器人程序320)。如框308所示，在所选文本被文本到语音转换合成器读出时，所选文本于是被自动处理以确定参数来添加一个或多个畸变。在框308中，在确定所述参数时可使用一个或多个语言模型。下面将更详细地讨论的这些畸变可包括谱频率翘曲、元音持续时间翘曲、音量翘曲、单词间时间的翘曲。

[0037] 一旦在框308中由文本到语音转换合成器产生所选文本的话音，该技术就在框310中创建音频质询。在音频质询的创建期间，可对框308中生成的话音 应用一个或多个附加畸变。这些畸变可以是添加回声、背景话音或音乐。在将背景音乐或话音添加到框308中生成的话音之前可将畸变应用于背景音乐或话音。可用与前景话音的生成类似的方式来生成背景话音，例如通过从一库中选择一文本句子或一组相关或不相关的单词，然后应用一语言模型以及文本到语音转换合成器来生成背景话音。当话音被文本到语音转换合成器生成时，可确定以及修改参数来应用一个或多个畸变。这些畸变可与前景话音的生成期间TTS合成器内部应用的畸变类似。背景话音可以具有不同的语言。在一个实施例中，添加的背景话音具有与框308中生成的前景话音的语言相同的语言。在另一实施例中，添加的背景话音具有与框308中生成的前景话音的语言不同的语言。在使用TTS合成器生成话音期间和之后添加失真有助于创建对人来说相对容易识别、但对计算机却难以识别的音频质询，并且在所生成的音频质询间引入变化。

[0038] 一旦在框310中生成了一音频质询，则下一步骤是将该音频质询发送并呈现给一未知用户以供标识，如框312所示。该未知用户然后被要求通过键入或说出对该音频质询的回答来进行响应，如框314所示。注意，即使在预期回答是所选文本串时，由于语音识别不能将说出的回答正确地转换成下一框中使用的文本回答，攻击者不能播放该音频HIP质询作为说出的回答。如框316所示，然后将该用户的回答与预期回答相比较。在一个实施例中，该用户的回答被说出来。在将说出的回答与预期回答比较之前，语音识别技术被应用来将说出的回答转换成文本回答。仅当键入的回答被确定为与预期回答相匹配时，才认为该未知用户是人类(框318)。否则，该未知用户被认为是机器人程序(框320)。在一个实施例中，要求匹配是精确的。在另一实施例中，匹配不必是精确的。可允许两个回答之间的某种失配。只要失配处于某一预定误差容限或阈值内，该用户的回答就仍被确定为与预期回答相匹配。

[0039] 在确定用户的回答是否与预期回答匹配时，在比较两个回答之前，框316中的验证器可以对各回答进行规范化以去除相同表达的变体。该规范化可去除无关紧要的字符或单词，以及用标准、等价的单词来替换一个或多个单词。例如，“I’m”可被“I am”替换，而“intl.”可被“international”替换。在又一实施例中，各回答可被转换成声音(即音素)的串，比较基于音素而非文本。

[0040] 在框316中可使用许多技术来计算两个回答之间的误差。在一个实施例中，编辑距离被用来计算两个文本或音素串之间的误差。上段中提及的规范化阶段可在计算编辑距离之前被应用。对于编辑距离的计算可基于单词或音素，或基于字符。当对单词计算误差时，如果一个单词是另一个的变体，例如另一单词的复数形式，或者两个单词之间的差异处于某一误差容限范围，则这两个单词可被认为是相同的。当对音素计算误差时，在计算两个回答的误差时，两个类似发音的音素可被认为是相同的。

[0041] 图4示出用于实践文本到语音转换音频HIP技术的另一实施例的又一示例性过程400。一般来说，在该实施例中，该技术通过首先定义离散文本句子及其预期回答的库来操作，这些预期回答要求用户理解句子的语义含义以便提供对句子的正确回答，如框402所示。在一个实施例中，文本是基于预设的一组规则自动生成的不重复的句子。库中的文本句子通常是指令或问题，对于该指令或问题，要求理解句子来提供正确的答案。例如，一组规则可产生与物品的增加或减少有关的许多问题，其中物品可以是任何常见对象，例如苹果、狗或飞机等。通过使用不同的数量和物品，可生成许多问题，例如“Simon ate three apples yesterday and has eaten two bananas today.What is the total number of fruits Simon has eaten since yesterday?”(“西蒙昨天吃了三个苹果，今天吃了两个香蕉。西蒙从昨天开始吃的水果总数是多少?”)可以改变主语、时间、数量以及物品名称来生成更多问题。另一组规则可使用乘法和/或除法、和/或加法和乘法来产生许多问题。作为另一示例，一组规则可通过要求用户以特定方式输入回答来生成问题，例如通过提供一句子然后要求用户以相反的顺序输入说出的单词的第二个字母，或输入第三个单词然后是其前一单词。该组规则也可生成许多问题。文本到语音转换音频HIP技术改变同一组规则所生成的句子的模式，以及对使用不同组规则生成的句子进行交织，来生成音频HIP质询，以便防止机器人程序基于生成音频HIP质询时使用的该组规则，或通过知道如何基于特定模式或关键词来提供正确的回答来正确地归类音频HIP。

[0042] 文本句子与它们的合适的答案或预期回答一起被存储。给定这一文本句子库，如框404所示，该技术从该库中自动选择一个或多个文本句子，用于创建要在确定未知的计算机用户是人类还是机器人程序时使用的音频质询。该所选 句子然后可被自动处理以确定在该所选句子被文本到语音转换合成器读出时可被添加的一个或多个畸变，如框406所示。下面将更详细地讨论的这些畸变包括谱频率翘曲、元音持续时间翘曲、音量翘曲、以及单词间时间的变化。在创建音频HIP时，可将诸如背景添加和回声添加之类的一个或多个附加畸变应用于由文本到语音转换合成器生成的话音，如框408所示。然而应注意，在一个实施例中，要求语义理解的句子在被文本到语音转换合成器读出之时或之后该句子不被畸变。未被畸变的音频HIP质询依赖于对该质询的语义理解来确定未知用户是人类还是机器人程序。语义理解防止机器人程序提供正确的回答。

[0043] 如框410所示，下一步骤是将该音频质询呈现给未知方来标识。然后该未知方被要

求要么通过键入要么通过说出合适的回答来对该要求语义理解的句子进行回答，如框412所示。通过应用语音识别技术，说出的回答可被转换成文本回答。回答可被转换成表示该回答如何被发音的音素串。可对回答应用规范化来用表达该回答的标准方式替换变体，无关紧要的字符或单词也可被去除。用户的回答然后与该音频质询的预期回答相比较，以确定它们是否匹配，如框414所示。仅当用户的回答被确定为与预期回答相匹配时，该未知用户才被认为是人类，如框416所示。否则，该未知用户被认为是机器人程序，如框418所示。在一个实施例中，仅当两个回答彼此精确匹配时才将该两个回答确定为彼此匹配。在另一实施例中，如果两个回答的误差处于容限范围内，则这两个回答被确定为彼此匹配。可使用不同的技术来计算两个回答的误差，例如编辑距离或其变体。对图3中所示的示例性过程描述的许多技术也可被应用于图4中所述的示例性过程。

[0044] 1.5该技术的各种实施例的细节。

[0045] 已经描述了用于实践文本到语音转换音频HIP技术的示例性架构和示例性过程，下面的段落提供了用于实现该技术的各种实施例的各种细节。

[0046] 1.5.1可被应用的各种畸变

[0047] 如上所讨论的，在从所选文本创建音频质询期间可应用一个或多个畸变。这些畸变可在生成文本的话音之时和/或之后被应用。这些畸变可随产生音频质询的每个实例而变化。在所选文本被文本到语音转换合成器读出时，和/或通过 在创建音频HIP质询时对所生成的话音进行背景添加和回声添加，文本到语音转换音频HIP技术可以采用谱频率翘曲、诸如元音持续时间翘曲之类的可发音的声音的变化、话音音量的变化、以及相邻单词间时间的变化。下文描述在创建用于确定未知用户是人类还是机器人程序的音频HIP质询时应用这些和其他畸变的细节。

[0048] 1.5.1.1谱频率翘曲

[0049] 在将所选文本转换成话音时可应用许多不同类型的频率翘曲来畸变所生成的话音，以便使机器人程序更难以识别该音频质询。例如，在音频质询的生成期间可应用一个或多个频率翘曲畸变来扭曲所生成的话音。为了这样做，各种翘曲函数和参数被确定并用于随时间以及随不同的音频质询来改变谱频率翘曲效应。

[0050] 在该文本到语音转换音频HIP技术的一个实施例中，为了执行谱频率翘曲，使用具有参数α的翘曲函数，α可随时间t改变。同时，使用函数$\hat{\omega}=\Psi_{\alpha(t)}(\omega)$来执行该变换。翘曲函数可以是线性的、分段线性的、双线性的或非线性的。在一个实施例中，此处所述的文本到语音转换音频技术使用基于具有单位增益的简单一阶全通滤波器的双线性频率翘曲函数：

[0051] $$\psi_{\alpha(t)}(z)=\frac{z^{-1}-\alpha(t)}{1-\alpha(t)z^{-1}}$$

[0052] 或

[0053] $$\Psi_{\alpha(t)}(\omega)=\omega+2\tan^{-1}\frac{\alpha(t)\sin(\omega)}{1-\alpha(t)\cos(\omega)}$$

[0054] 在一个实施例中，翘曲参数α(t)最好随时间平滑改变。从而，这里使用一正弦函数，如下：

[0055] α(t)＝B+A sin((k+t)/T*2*π)

[0056] 其中A、B和T是翘曲范围、翘曲中心和翘曲周期，并且要么被手动地设置要么在特定范围内变化，k是初始相位并且要么被随机地要么被非随机地设为[0,T-1]内的某一值。

[0057] 应当注意，上述翘曲函数是可与此处所述的技术一起被使用的一个示例性的翘曲函数。可以使用各种其他翘曲函数，这些其他翘曲函数或它们的参数也可以随时间变化或者可以随时间被平滑地应用。

[0058] 1.5.1.2元音持续时间翘曲

[0059] 在该文本到语音转换音频HIP技术的一个实施例中，当文本被文本到语音转换合成器读出时，改变可发音的声音的发音的持续时间以便对所选文本串的所生成的话音进行畸变。例如，在一个实施例中，在由文本到语音转换合成器读出所选文本时，使用元音持续时间翘曲来改变元音的发音的持续时间。在采用元音持续时间翘曲的该实施例中，文本到语音转换音频HIP技术首先对每个元音设置仍能被人认知的最小持续时间和最大持续时间，然后在文本到语音转换合成器生成所选文本的话音期间随机调整元音持续时间。还应注意，以类似的方式也可改变特定辅音。

[0060] 1.5.1.3音量翘曲

[0061] 也可应用音量翘曲来在文本到语音转换合成器读出所选文本时改变可发音的声音的音量。在一个实施例中，设置最小音量和最大音量，对某一发音应用最小音量和最大音量之间的某一随机音量来应用音量翘曲。

[0062] 1.5.1.4改变单词间的持续时间

[0063] 在所选文本被文本到语音转换合成器读出时，两个单词之间的持续时间也可被改变。在一个实施例中，设置最小持续时间和最大持续时间，可随机选择最小持续时间和最大持续时间之间的某一持续时间并应用于两个相邻单词的持续时间。如果所选持续时间是负的，则两个相邻单词用指定的重叠来发音。单词之间的持续时间的这个变化使得ASR系统难以将句子分段成个体单词。

[0064] 1.5.1.5背景和回声添加

[0065] 文本到语音转换音频HIP技术还可向所选文本的生成的话音添加一个或多个畸变。在某些实施例中，可将背景和回声应用于文本到语音转换合成器读出的话音。例如，背景可以是噪声、音乐、相同或其他语言的语音话音等。回声也可被添加到所选文本的所生成的话音。例如，可以随机设定衰减百分比、延迟时间以及初始回声音量。此外，在生成所选文本串的话音之后应用的一个或多个畸变可包括将通过某一文本到语音转换技术生成的另一语音添加到该文本串的话音的背景，来创建音频质询。在一个实施例中，添加到背景的该附加的语音可以是与所选文本串的语言不同语言的语音。可以将背景语音选择成所生成的音频质询的大多数目标人群不知道的一种语言。人类可以容易地标识不 同语言的语音，并聚焦于人类用户知道的前景语音。机器人程序可能难以将前景语音和背景语音相区分，从而不能识别前景语音。在另一实施例中，背景语音可具有与前景语音相同的语言。可通过用一文本到语音转换合成器读出一句子或一组相关或不相关的单词来生成背景语音。背景语音的音量可在合适的范围内变化，使得前景语音能够容易地被人类标识。在添加背景时可应用一个或多个畸变。例如，在背景语音被一文本到语音转换合成器读出之时或之后可将一个或多个畸变应用于添加的背景语音。这些畸变可包括但不限于：频率翘曲、可发音的声音的持续时间翘曲、音量翘曲以及单词间的持续时间的变化。一个或多个畸变可被应用于

由一文本到语音转换合成器生成的背景语音。例如，在将背景语音添加到前景语音之前，可将回声添加到背景语音。此外，背景语音可以采用无意义的语音或录制的音频的形式。在背景语音具有与前景语音相同的语言的实施例中，无意义的背景语音可以帮助人类标识和识别前景语音。

[0066] 1.5.2音频HIP质询中使用的文本

[0067] 在该文本到语音转换音频HIP技术的某些实施例中，每个音频HIP质询都是经一文本到语音转换合成器说出的句子。该文本到语音转换音频HIP技术的一个简单实施例从一文章中通常在特定范围内选择一具有合适单词长度的句子，并使用该文本到语音转换合成器来说出所选句子。在其他实施例中，音频HIP质询是经一文本到语音转换合成器说出的相关或不相关单词的串。这些单词可从单词库中选择，单词库是通过去除人类在识别时可造成人类混淆的单词以及无关紧要的单词而从字典中构建的。

[0068] 该技术向未知用户呈现音频质询，并要求该未知用户键入或说出对该音频质询的回答。在某些实施例中，该未知用户被要求用他或她所听到的句子或单词串来回答。这通常在不需要对所选文本的语义理解的情况下被使用。该未知用户仅需要正确地识别说出的句子或单词串。这些实施例具有能够容易地生成不同语言的音频HIP质询的优点。在其他实施例中，该未知用户需要理解该说出的句子以提供正确的回答。该句子通常是用一组或多组规则自动生成的指令或问题。这些实施例具有在所生成的音频质询中应用附加的安全等级的优点。未知用户不仅需要正确地识别说出的句子，还要正确地理解该句子以便提供正确的回答。当需要语义理解来回答音频质询时，预期的答案通常随该句子被生 成，并与该句子一起被存储在库中。

[0069] 1.5.3语义理解

[0070] 尽管该技术生成的上述音频HIP质询中的许多都不需要对用作音频质询的句子的语义理解，但是在该文本到语音转换音频HIP技术的某些实施例中，可添加附加的机制来帮助将人类和机器人程序区分开来。在该情况下，要求对音频HIP质询的句子的理解以便通过测试。该句子可以是问题或指令。例如，在某些实施例中，该技术基于问题或指令的类型定义了多个类别的问题或指令。一个或多个规则可与每一类别相关联来帮助自动生成文本句子及其预期回答。需要对这种句子的语义理解来提供正确的回答。人类理解该句子，因此能容易地提供正确的答案。另一方面，机器人程序不具有理解该句子的能力，因此不能提供正确的答案。因此，该句子本身就是一HIP质询。如果该句子被用作所选文本来生成音频质询，即使机器人程序正确地识别出该音频质询的文本，它们仍然不能提供正确的回答并通过HIP测试，因为它们不理解该句子的语义含义。额外类型的问题和指令可被添加到该系统。在一个实施例中，一种类别是预期回答是基于句子的特定一串字符或单词。例如，它可以是随机选择的句子后跟一指令，该指令要求用户输入前一句子中单词的第二个字母，或以逆序输入最后两个单词等等。与该类别相关联的该组规则确定不同类型的指令(因此确定用于同一所选句子的不同预期回答)以及用来陈述产生相同预期回答的等价指令的不同方式。由于机器人程序不理解该指令，因此它们将不能提供正确的回答。一旦这一合成句子(随机选择的句子加上后跟的指令)被生成，预期回答也被生成。该一个或多个预期回答可被添加到库中，将以后在生成音频HIP质询时被选择。在另一实施例中，当预期回答是特定的计算结果时要使用一种类别。例如，与该类别相关联的该组规则是生成与计算结果有关

的不同问题以及产生相同计算结果的不同表达方式。例如，生成的句子可以是："Simon ate three apples yesterday and has eaten two bananas today,which day did he eat more fruits in terms of units of fruit?"("西蒙昨天吃了三个苹果并且今天已经吃了两根香蕉，以水果数量来计的话他哪天吃了更多的水果?")该句子的预期答案也被自动生成。通过改变主语、时间、要问的问题、以及陈述同一件事的等价方式，该技术能生成多个句子及其预期回答。

[0071] 2.0示例性操作环境：

[0072] 本文所述的文本到语音转换音频HIP技术可在多种类型的通用或专用计算系统环境或配置内操作。图5示出可在其上实现本文所述的文本到语音转换音频HIP技术的各种实施例和各元素的通用计算机系统的简化示例。应当注意，图5中由折线或虚线所表示的任何框表示简化计算设备的替换实施方式，并且以下描述的这些替换实施方式中的任一个或全部可以结合贯穿本文所描述的其他替换实施方式来使用。

[0073] 例如，图5示出了总系统图，其示出简化计算设备500。这样的计算设备通常可以在具有至少一些最小计算能力的设备中找到，这些设备包括但不限于个人计算机、服务器计算机、手持式计算设备、膝上型或移动计算机、诸如蜂窝电话和PDA等通信设备、多处理器系统、基于微处理器的系统、机顶盒、可编程消费电子产品、网络PC、小型计算机、大型计算机、音频或视频媒体播放器等。

[0074] 为允许设备实现文本到语音转换音频HIP技术，该设备应当具有足够的计算能力和系统存储器以实现基本的计算操作。具体而言，如图5所示，计算能力一般由一个或多个处理单元510示出，并且还可包括一个或多个GPU 515，这两者中的任一种或全部与系统存储器520通信。注意，通用计算设备的处理单元510可以是专用微处理器，如DSP、VLIW、或其他微控制器、或可以是具有一个或多个处理核的常规CPU，包括多核CPU中的基于GPU专用核。

[0075] 另外，图5的简化计算设备还可包括其他组件，诸如例如通信接口530。图5的简化计算设备还可包括一个或多个常规计算机输入设备540(例如，定点设备、键盘、音频输入设备、视频输入设备、触觉输入设备、用于接收有线或无线数据传输的设备等)。图5的简化计算设备还可包括其他光学组件，诸如例如一个或多个常规计算机输出设备550(例如，显示设备555、音频输出设备、视频输出设备、用于传送有线或无线数据传输的设备等)。注意，通用计算机的典型的通信接口530、输入设备540、输出设备550、以及存储设备550对本领域技术人员而言是公知的，并且在此不会详细描述。

[0076] 图5的简化计算设备还可包括各种计算机可读介质。计算机可读介质可以是可由计算机500经由存储设备560访问的任何可用介质，并且包括是可移动 570和/或不可移动580的易失性和非易失性介质，该介质用于存储诸如计算机可读或计算机可执行指令、数据结构、程序模块或其他数据等信息。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但不限于：计算机或机器可读介质或存储设备，诸如DVD、CD、软盘、磁带驱动器、硬盘驱动器、光盘驱动器、固态存储器设备、RAM、ROM、EEPROM、闪存或其他存储器技术、磁带盒、磁带、磁盘存储或其他磁存储设备、或可用于存储所需信息并且可由一个或多个计算设备访问的任何其他设备。

[0077] 诸如计算机可读或计算机可执行指令、数据结构、程序模块等信息的存储还可通

过使用各种上述通信介质中的任一种来编码一个或多个已调制数据信号或载波或其他传输机制或通信协议来实现，并且包括任何有线或无线信息传递机制。注意，术语“已调制数据信号”或“载波”一般指以对信号中的信息进行编码的方式设置或改变其一个或多个特征的信号。例如，通信介质包括诸如有线网络或直接线连接等携带一个或多个已调制数据信号的有线介质，以及诸如声学、RF、红外线、激光和其他无线介质等用于传送和/或接收一个或多个已调制数据信号或载波的无线介质。上述通信介质的任一组合也应包括在通信介质的范围之内。

[0078] 此外，可以按计算机可执行指令或其他数据结构的形式存储、接收、传送或者从计算机或机器可读介质或存储设备和通信介质的任何所需组合中读取体现此处所描述的文本到语音转换音频HIP技术的部分或全部的软件、程序和/或计算机程序产品或其各部分。

[0079] 最后，本文描述的文本到语音转换音频HIP技术还可在由计算设备执行的诸如程序模块等计算机可执行指令的一般上下文中描述。一般而言，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等。本文描述的各实施例还可以在其中任务由通过一个或多个通信网络链接的一个或多个远程处理设备执行或者在该一个或多个设备的云中执行的分布式计算环境中实现。在分布式计算环境中，程序模块可以位于包括媒体存储设备在内的本地和远程计算机存储介质中。此外，上述指令可以部分地或整体地作为可以包括或不包括处理器的硬件逻辑电路来实现。

[0080] 还应当注意，可以按所需的任何组合来使用此处所述的上述替换实施例的任一个或全部以形成另外的混合实施例。尽管用结构特征和/或方法动作专用的语言描述了本主题，但可以理解，所附权利要求书中定义的主题不必限于上述具体特征或动作。上述具体特征和动作是作为实现权利要求的示例形式公开的。

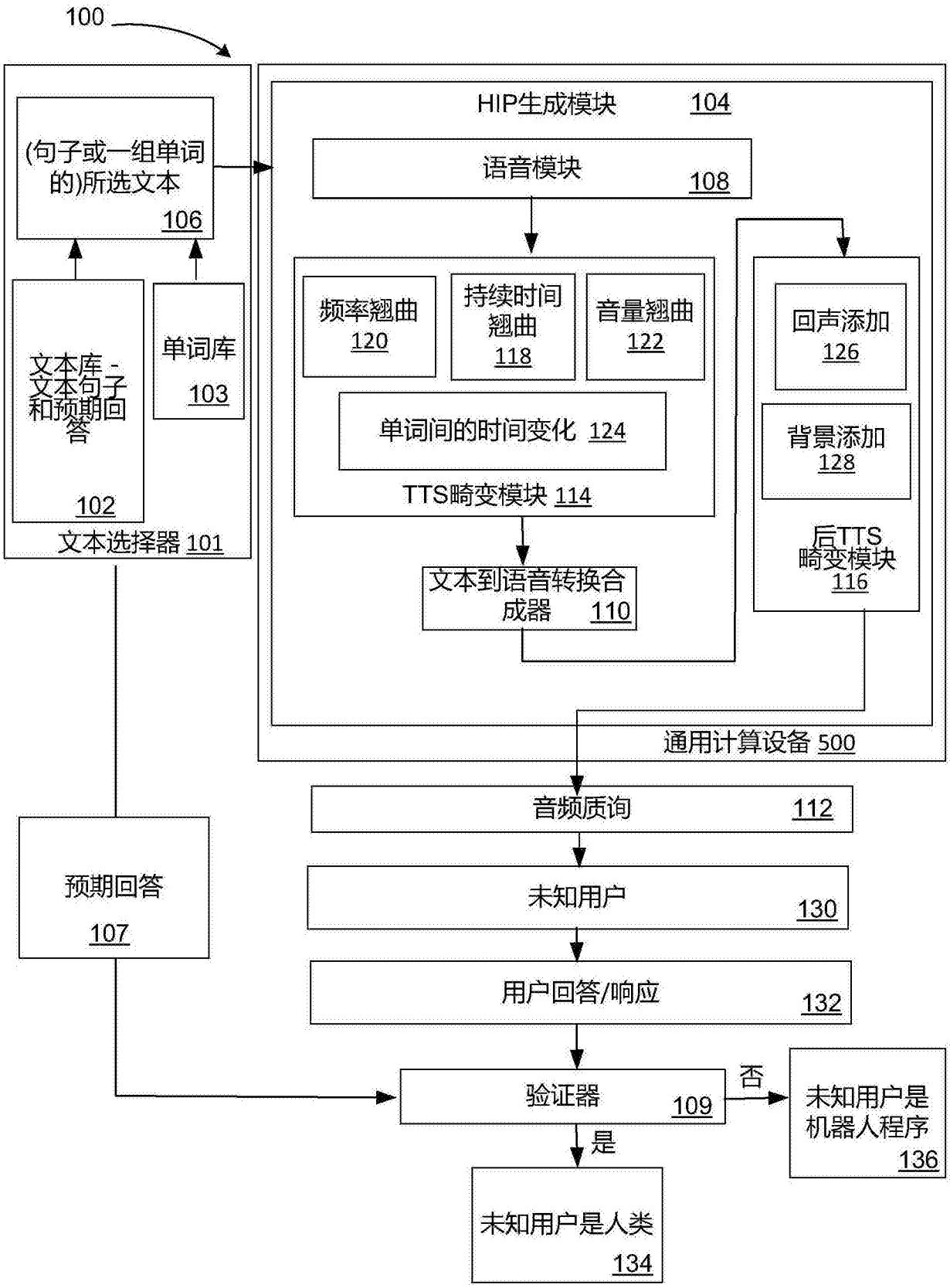

100
(句子或一组单词的)所选文本
106
文本库 - 文本句子和预期回答
102
单词库
103
文本选择器 101
HIP生成模块 104
语音模块 108
频率翘曲 120
持续时间翘曲 118
音量翘曲 122
单词间的时间变化 124
TTS畸变模块 114
文本到语音转换合成器 110
回声添加 126
背景添加 128
后TTS畸变模块 116
通用计算设备 500
预期回答 107
音频质询 112
未知用户 130
用户回答/响应 132
验证器 109
否
未知用户是机器人程序 136
是
未知用户是人类 134

图1

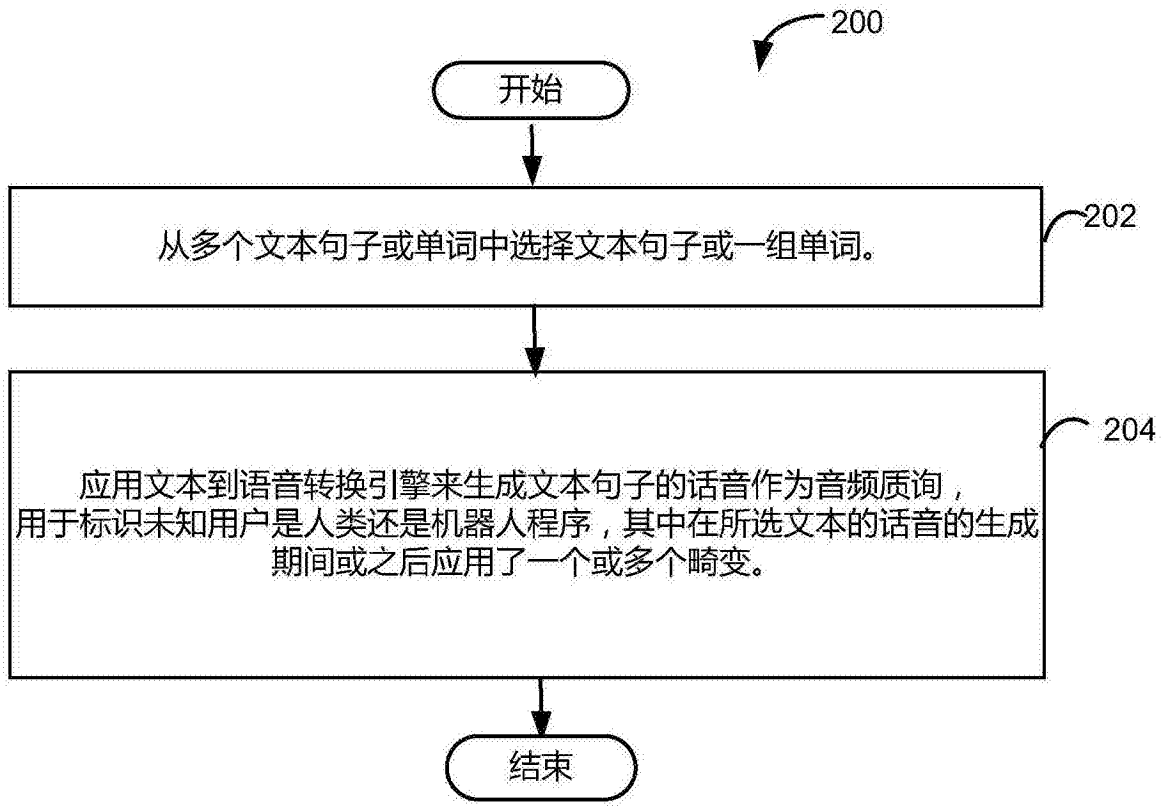

200
开始
从多个文本句子或单词中选择文本句子或一组单词。
202
204
应用文本到语音转换引擎来生成文本句子的话音作为音频质询，
用于标识未知用户是人类还是机器人程序，其中在所选文本的话音的生成
期间或之后应用了一个或多个畸变。
结束

图2

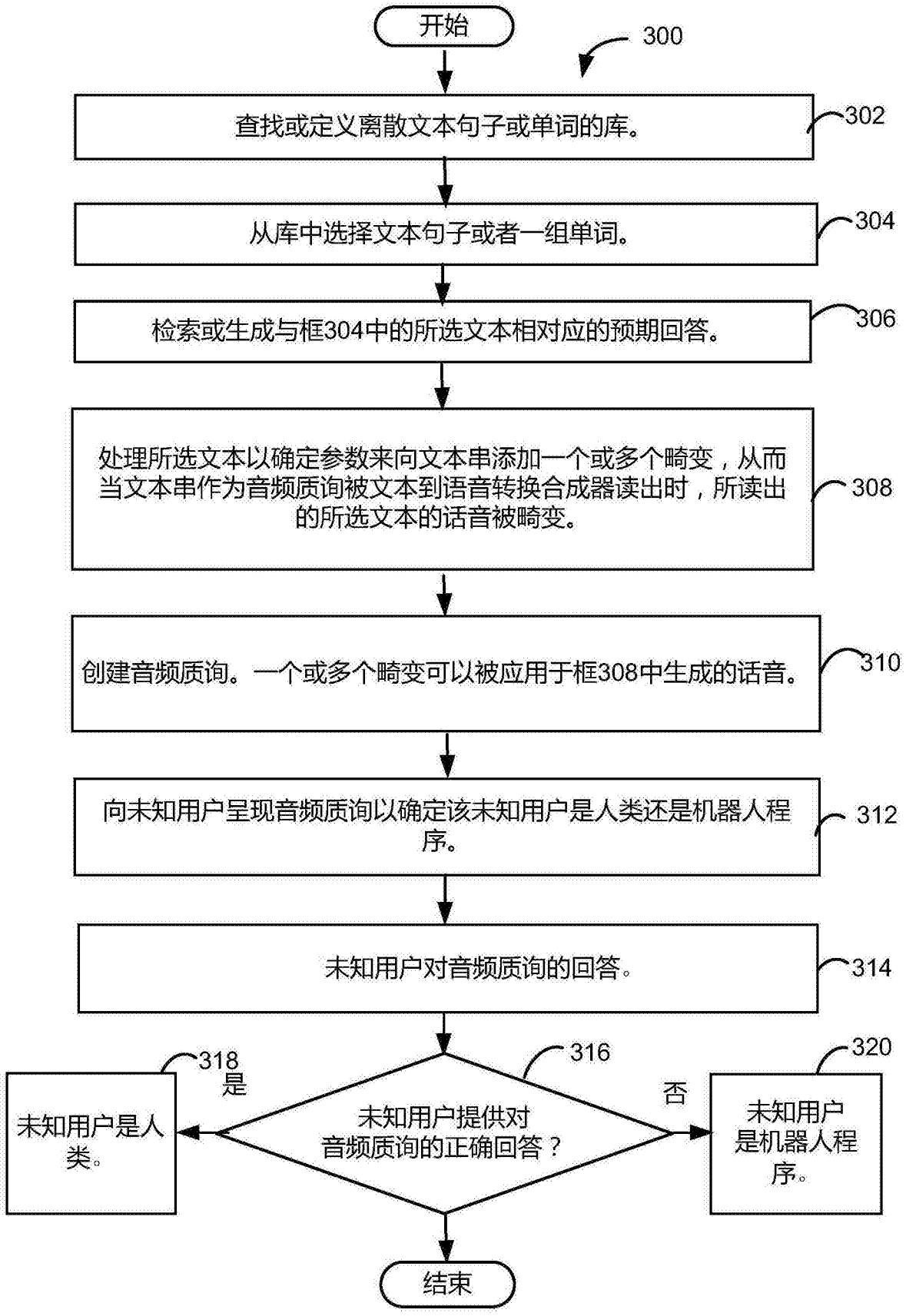

开始
300
查找或定义离散文本句子或单词的库。
302
从库中选择文本句子或者一组单词。
304
检索或生成与框304中的所选文本相对应的预期回答。
306
处理所选文本以确定参数来向文本串添加一个或多个畸变，从而当文本串作为音频质询被文本到语音转换合成器读出时，所读出的所选文本的话音被畸变。
308
创建音频质询。一个或多个畸变可以被应用于框308中生成的话音。
310
向未知用户呈现音频质询以确定该未知用户是人类还是机器人程序。
312
未知用户对音频质询的回答。
314
未知用户提供对音频质询的正确回答？
316
是
否
318
未知用户是人类。
320
未知用户是机器人程序。
结束

图3

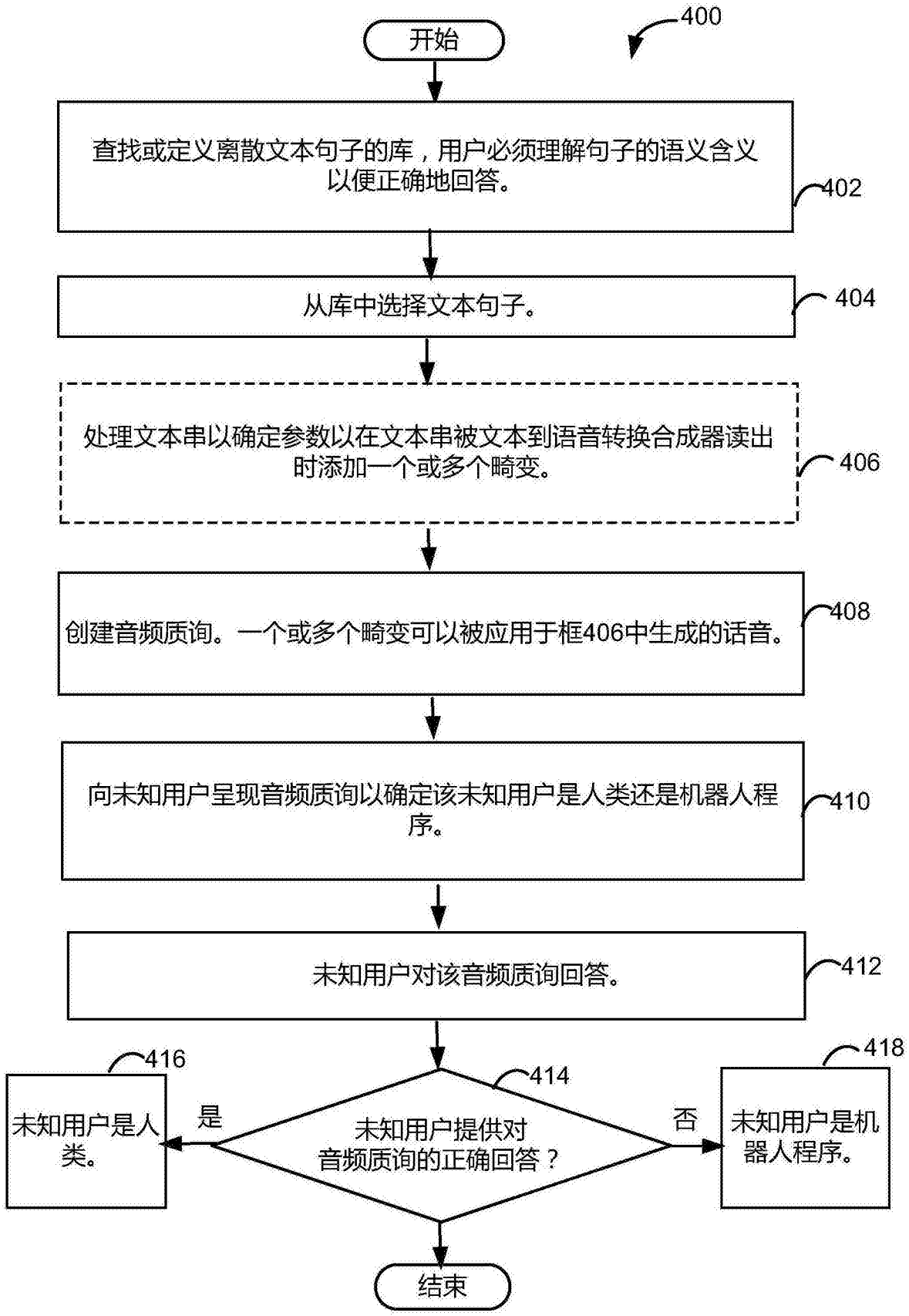

400
开始
查找或定义离散文本句子的库，用户必须理解句子的语义含义以便正确地回答。
402
从库中选择文本句子。
404
处理文本串以确定参数以在文本串被文本到语音转换合成器读出时添加一个或多个畸变。
406
创建音频质询。一个或多个畸变可以被应用于框406中生成的话音。
408
向未知用户呈现音频质询以确定该未知用户是人类还是机器人程序。
410
未知用户对该音频质询回答。
412
未知用户提供对音频质询的正确回答？
414
是
否
未知用户是人类。
416
未知用户是机器人程序。
418
结束

图4

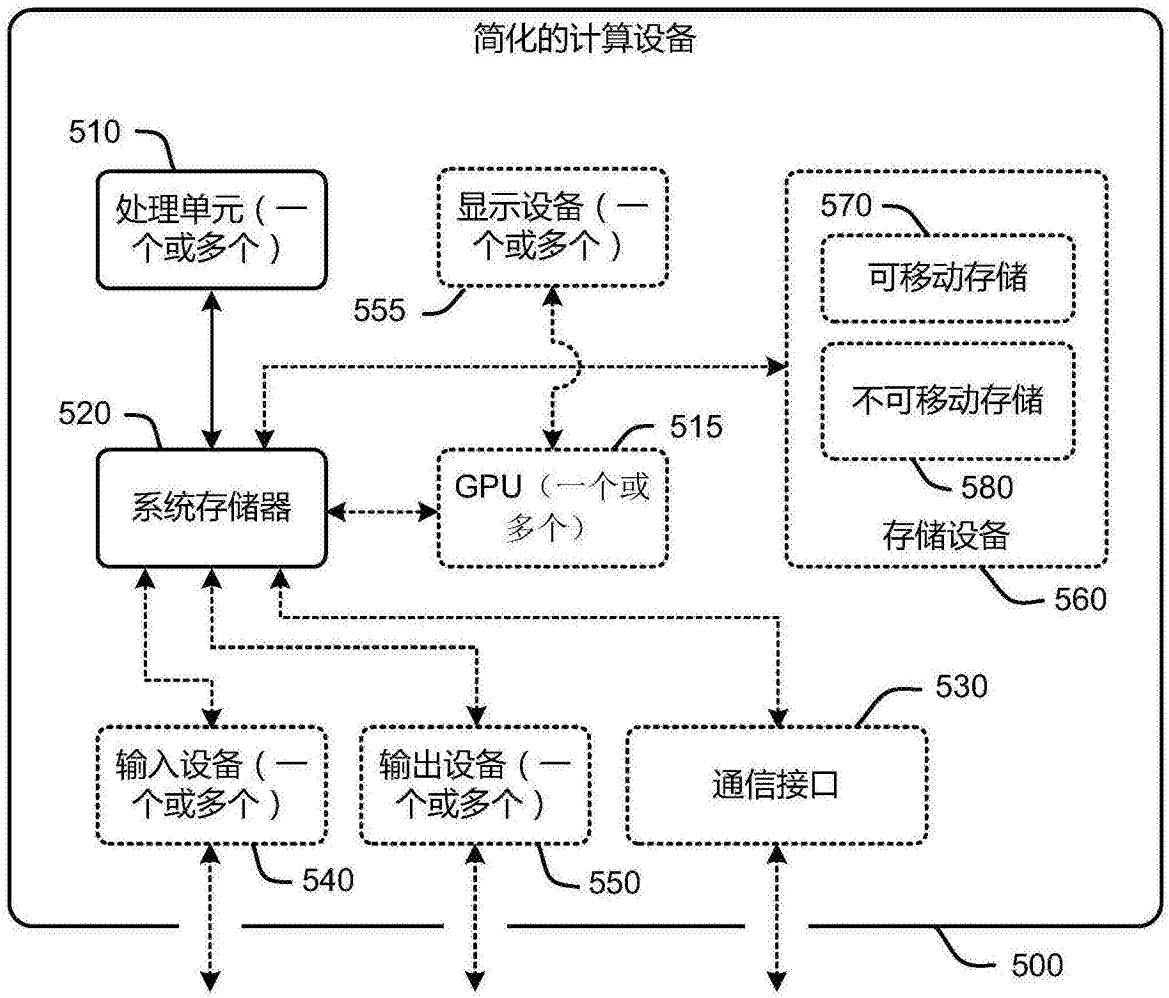

简化的计算设备
510
处理单元（一个或多个）
显示设备（一个或多个）
555
570
可移动存储
不可移动存储
580
存储设备
560
520
系统存储器
515
GPU（一个或多个）
530
输入设备（一个或多个）
输出设备（一个或多个）
通信接口
540
550
500

图5