

[19] 中华人民共和国国家知识产权局



[12] 发明专利申请公布说明书

[21] 申请号 200480010899.X

[51] Int. Cl.

G10L 13/00 (2006.01)

G10L 13/08 (2006.01)

G10L 13/02 (2006.01)

G10L 13/06 (2006.01)

[43] 公开日 2007 年 1 月 10 日

[11] 公开号 CN 1894739A

[22] 申请日 2004.4.28

[21] 申请号 200480010899.X

[30] 优先权

[32] 2003.5.9 [33] US [31] 10/434,683

[86] 国际申请 PCT/US2004/013366 2004.4.28

[87] 国际公布 WO2004/100638 英 2004.11.25

[85] 进入国家阶段日期 2005.10.24

[71] 申请人 思科技术公司

地址 美国加利福尼亚州

[72] 发明人 尼克拉斯·J·卡塔艾

[74] 专利代理机构 北京东方亿思知识产权代理有限公司

代理人 王 怡

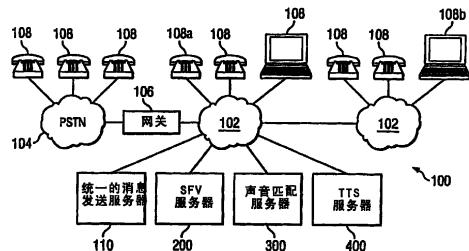
权利要求书 5 页 说明书 14 页 附图 4 页

[54] 发明名称

依赖于源的文本到语音系统

[57] 摘要

本发明提供了一种从文本消息生成语音的方法，该方法包括确定与文本消息的源相关联的声音的语音特征向量，并且比较该语音特征向量和多个语者模型。该方法还包括基于比较选择出语者模型之一作为声音的优选匹配，并且基于所选出的语者模型从文本消息生成语音。



1. 一种从文本消息生成语音的方法，包括以下步骤：
确定与文本消息的源相关联的声音的语音特征向量；
比较所述语音特征向量和多个语者模型；
基于所述比较选择出所述语者模型之一作为所述声音的优选匹配；和
基于所选出的语者模型从所述文本消息生成语音。
2. 如权利要求 1 所述的方法，其中，所述确定步骤包括：
接收所述声音的采样；和
分析所述采样来确定所述声音的语音特征向量。
3. 如权利要求 1 所述的方法，其中，所述确定步骤包括：
请求作为所述文本消息的源的端点提供所述语音特征向量；和
从所述端点接收所述语音特征向量。
4. 如权利要求 1 所述的方法，其中，所述生成步骤包括将生成所述语音的命令传输到文本到语音服务器，所述命令包括所选出的语者模型，其中所述文本到语音服务器基于所选出的语者模型生成所述语音。
5. 如权利要求 1 所述的方法，其中：
所述语音特征向量包括用于高斯混合模型的特征向量；并且
所述比较步骤包括将与所述语音特征向量相关联的第一高斯混合模型与多个第二高斯混合模型相比较，其中每个第二高斯混合模型与所述语者模型中的至少一个相关联。
6. 如权利要求 1 所述的方法，还包括：
生成多个模型声音采样；和
分析所述模型声音采样来确定所述每个模型声音采样的语者模型。
7. 如权利要求 6 所述的方法，其中，所述模型声音采样是基于与所述声音采样相关联的文本采样生成的。
8. 如权利要求 1 所述的方法，其中，所述方法的步骤由通信网络中的端点实现。
9. 如权利要求 1 所述的方法，其中，所述方法的步骤在通信网络中的

声音匹配服务器中实现。

10. 如权利要求 1 所述的方法，其中：

所述方法的步骤在统一的消息发送系统中实现；并且
所述语音特征向量在用户概况中被关联到提供所述文本消息的用户。

11. 一种声音匹配服务器，包括：

接口，可操作来执行下面的功能：

接收与文本消息的源相关联的声音的语音特征向量；以及
传输命令到文本到语音服务器，指示所述文本到语音服务器基
于所选出的语者模型从所述文本消息生成语音；和

处理器，可操作来执行下面的功能：

比较所述语音特征向量和多个语者模型；以及
基于所述比较选择所述语者模型之一作为所述声音的优选匹
配。

12. 如权利要求 11 所述的服务器，还包括存储器，其可操作来存储所
述多个语者模型。

13. 如权利要求 11 所述的服务器，其中：

所述接口还可操作来使所述文本到语音服务器生成多个模型声音采
样；并且

所述语者模型是基于所述模型声音采样的分析而确定的。

14. 如权利要求 13 所述的服务器，其中，所述模型声音采样是基于与
所述声音采样相关联的文本采样而生成的。

15. 如权利要求 11 所述的服务器，其中：

所述接口还可操作来向作为所述文本消息的源的端点传输请求所述语
音特征向量的请求；并且

所述接口从所述端点接收所述语音特征向量。

16. 如权利要求 11 所述的服务器，其中：

所述语音特征向量包括用于高斯混合模型的特征向量；并且

所述比较步骤包括将与所述语音特征向量相关联的第一高斯混合模型
与多个第二高斯混合模型相比较，其中每个第二高斯混合模型与所述语者

模型中的至少一个相关联。

17. 如权利要求 11 所述的服务器，其中：

所述服务器是统一的消息发送系统的一部分；并且

所述语音特征向量在用户概况中被关联到提供所述文本消息的用户。

18. 一种端点，包括：

第一接口，可操作来从源接收文本消息；和

处理器，可操作来执行下面的功能：

确定与文本消息的源相关联的声音的语音特征向量；

比较所述语音特征向量和多个语者模型；

基于所述比较选择出所述语者模型之一作为所述声音的优选匹配；以及

基于所选出的语者模型从所述文本消息生成语音；和

第二接口，可操作来向用户输出所生成的语音。

19. 如权利要求 18 所述的端点，其中，所述第一接口还可操作来执行下面的功能：

向所述文本消息的源传输请求所述语音特征向量的请求；和

接收响应于所述请求的所述语音特征向量。

20. 如权利要求 18 所述的端点，其中：

所述第一接口还可操作来从所述文本消息的源接收声音采样；并且

所述处理器还可操作来分析所述声音采样以确定所述语音特征向量。

21. 如权利要求 18 所述的端点，其中：

所述第一接口还可操作来从所述文本消息的源接收语音；

所述第二接口还可操作来输出所接收到的语音；并且

所述处理器还可操作来分析所接收到的语音以确定所述语音特征向量。

22. 一种系统，包括：

声音匹配服务器，可操作来执行下面的功能：

比较语音特征向量和多个语者模型；以及

基于所述比较选择出所述语者模型之一作为所述声音的优选匹

配；和

文本到语音服务器，可操作来基于所选出的语者模型从文本消息生成语音。

23. 如权利要求 22 所述的系统，还包括语音特征向量服务器，可操作来执行下面的功能：

接收语音；和

基于所述语音确定关联的语音特征向量，其中由所述声音匹配服务器比较的语音特征向量是从所述语音特征向量服务器接收到的。

24. 如权利要求 22 所述的系统，其中，所述声音匹配服务器还可操作来从所述语音特征向量服务器接收所述语者模型。

25. 如权利要求 24 所述的系统，其中：

所述声音匹配服务器还可操作来使所述文本到语音服务器生成多个模型声音采样；并且

所述语音特征向量服务器还可操作来分析所述声音采样以确定所述语者模型。

26. 如权利要求 22 所述的系统，其中：

所述文本到语音服务器是多个文本到语音服务器中的一个，每个文本到语音服务器可操作来使用不同的语者模型生成语音；并且

所述声音匹配服务器还可操作来基于哪个文本到语音服务器使用所选出的语者模型来选择所述文本到语音服务器之一以生成语音。

27. 包含在计算机可读介质中的软件，所述软件可操作来执行下面的步骤：

确定与文本消息的源相关联的声音的语音特征向量；

比较所述语音特征向量和多个语者模型；

基于所述比较选择出所述语者模型之一作为所述声音的优选匹配；和

基于所选出的语者模型从所述文本消息生成语音。

28. 如权利要求 27 所述的软件，其中，所述确定步骤包括：

接收所述声音的采样；和

分析所述采样以确定所述声音的语音特征向量。

29. 如权利要求 27 所述的软件，其中，所述确定步骤包括：
请求作为所述文本消息的源的端点提供所述语音特征向量；和
从所述端点接收所述语音特征向量。

30. 如权利要求 27 所述的软件，还可操作来执行下面的步骤：
生成多个模型声音采样；和
分析所述模型声音采样以确定所述每个模型声音采样的语者模型。

31. 一种系统，包括：
用于确定与文本消息的源相关联的声音的语音特征向量的装置；
用于比较所述语音特征向量和多个语者模型的装置；
用于基于所述比较选择出所述语者模型之一作为所述声音的优选匹配的装置；和
用于基于所选出的语者模型从所述文本消息生成语音的装置。

32. 如权利要求 31 所述的系统，其中所述用于确定的装置包括：
用于接收所述声音的采样的装置；和
用于分析所述采样以确定所述声音的语音特征向量的装置。

33. 如权利要求 31 所述的系统，其中所述用于确定的装置包括：
用于请求作为所述文本消息的源的端点提供所述语音特征向量的装置；和
用于从所述端点接收所述语音特征向量的装置。

34. 如权利要求 31 所述的系统，还包括：
用于生成多个模型声音采样的装置；和
用于分析所述模型声音采样以确定所述每个模型声音采样的语者模型的装置。

依赖于源的文本到语音系统

技术领域

本发明一般地涉及文本到语音系统，更具体地说，本发明涉及依赖于源的文本到语音系统。

背景技术

文本到语音（TTS）系统在电信网络中提供了多功能性。TTS 系统从诸如电子邮件、即时消息或者其他适当的文本之类的文本消息产生可听语音。TTS 系统的一个缺点是 TTS 系统所产生的声音常常是通用的，而未与提供该消息的具体源相关联。例如，文本到语音系统可以产生男声，而不管发送该消息的人是谁，导致难以判断出特定的消息是来自男性还是女性。

发明内容

根据本发明，文本到语音系统以与提供文本消息的人类似的声音方式提供了依赖于源的文本消息表现。这增强了 TTS 系统用户的能力，使其能够通过将消息与特定声音的发声相关联，从而确定文本消息的源。具体地说，本发明的某些实施例提供了依赖于源的 TTS 系统。

根据本发明的一个实施例，提供了一种从文本消息生成语音的方法，该方法包括确定与文本消息的源相关联的声音的语音特征向量，并且比较该语音特征向量和多个语者模型。该方法还包括基于比较选择出语者模型之一作为该声音的优选匹配，并且基于所选出的语者模型从文本消息生成语音。

根据本发明的另一个实施例，提供了一种声音匹配服务器，该服务器包括接口和处理器。该接口接收与文本消息的源相关联的声音的语音特征向量。该处理器比较该语音特征向量和多个语者模型，并且基于比较选择

语者模型之一作为声音的优选匹配。然后，接口传输命令到文本到语音服务器，指示该文本到语音服务器基于所选出的语者模型从文本消息生成语音。

根据本发明另一个实施例，提供了一种端点，该端点包括第一接口、第二接口和处理器。第一接口从源接收文本消息。处理器确定与文本消息的源相关联的声音的语音特征向量，比较该语音特征向量和多个语者模型，基于比较选择出语者模型之一作为声音的优选匹配，并且基于所选出的语者模型从文本消息生成语音。第二接口向用户输出所生成的语音。

本发明的某些实施例的重要的技术优点包括再现的语音，这种再现的语音更忠于原来提供消息的真人的语音。这向 TTS 系统的用户提供了次要线索，其增强了用户的识别消息源的能力，并且还在 TTS 接口中提供了更多的舒适性和灵活性。这也增加了 TTS 系统的满意度和有用性。

本发明的某些实施例的其他重要的技术优点包括 TTS 系统的互操作能力。在某些实施例中，TTS 系统可以从另一个可能未使用相同的 TTS 标记参数和语音生成方法的 TTS 系统接收信息。但是，即使这些系统不共享 TTS 标记参数和语音生成方法，该 TTS 系统也仍然可以从远程 TTS 系统接收语音信息。这允许这种实施例的特征适用于与不包括相同特征的其他 TTS 系统一起工作。

从下面所包括的附图、说明书和权利要求书，本发明的其他技术优点将对本领域的技术人员变清楚。此外，尽管上面已列举出了本发明的特定优点，但是各种实施例可以包括所列举出的优点的全部、某些、或者不包括这些优点。

附图说明

为了更全面地理解本发明及其优点，现在结合附图参考下面的描述，在附图中：

图 1 是根据本发明特定实施例的提供依赖于源的文本到语音的电信系统；

图 2 示出了在图 1 的网络中的语音特征向量服务器；

图 3 示出了在图 1 的网络中的声音匹配服务器；

图 4 示出了在图 1 的网络中的文本到语音服务器；

图 5 示出了根据本发明特定实施例的提供依赖于源的文本到语音的端点；以及

图 6 是示出了图 1 的网络的工作方法的一个示例的流程图。

具体实施方式

图 1 示出了电信网络 100，该电信网络允许端点 108 彼此交换文本和/或语音形式的消息。一般来说，网络 100 的组件实现用于从文本消息生成声音消息的技术，以使该声音消息的声学特征对应于与该文本消息的源相关联的声音的声学特征。在所示实施例中，网络 100 包括利用网关 106 髓合到公共交换电话网络（PSTN）104 的数据网络 102。耦合到网络 102 和 104 的端点 108 向用户提供通信服务。网络 100 中的各种服务器向端点 108 提供服务。具体地说，网络 100 包括语音特征向量（SFV）服务器 200、声音匹配服务器 300、文本到语音（TTS）服务器 400 和统一的消息发送服务器 110。在替换实施例中，由这些各种组件提供的功能和服务可被积聚在不同的或其他的组件内，或者分布在不同的或其他的组件之间，例如包括将服务器 200、300 和 400 集成到单个服务器，或者提供分布式体系结构，在该结构中，端点 108 执行所述服务器 200、300 和 400 的功能。

总地来说，网络 100 应用各种模式识别技术来确定与文本消息的源相关联的声音和可由 TTS 系统产生的数种不同的声音之一之间的最优匹配。一般来说，模式识别目的在于基于现有知识或从源数据的模式抽取的统计信息来对从源生成的数据进行分类。要被分类的模式通常是度量或观测量的组，它们定义适当的多维空间中的多个点。模式识别系统一般包括收集观测量的传感器、从观测量计算数值或符号信息的特征抽取机制、对观测量进行分类的分类方案、以及根据所抽取的特征描述观测量的描述方案。分类和描述方案可以是基于可用模式的，通常使用统计、句法或者神经分析方法已对这些可用模式作了分类或描述。统计方法基于概率系统生成的模式的统计特性；句法方法基于特征的结构相互关系；而神经方法采用在

神经网络中使用的神经计算程序。

网络 100 通过计算语音特征向量，从而将模式识别技术应用到声音。如同在下面的描述中所使用的，“语音特征向量”指描述语音的许多数学量中的任何一个。开始，网络 100 针对可由 TTS 系统生成的某一范围内的声音计算语音特征向量，并且将每一声音的语音特征向量关联到生成该声音所使用的 TTS 系统的设置。在下面的描述中，TTS 系统的这种设置被称作“TTS 标记参数”。一旦学会了 TTS 系统的声音，网络 100 就使用模式识别来比较新声音与所存储的声音。这些声音之间的比较可以包括数值值的基本比较，或者可包括更复杂的技术，例如假设检验，在这些比较中，声音识别系统使用数种技术中的任何一种来识别所考虑的声音的可能匹配，并且计算该声音匹配的概率分值。此外，诸如梯度下降或共轭梯度下降之类的优化技术可被用来选择候选者。使用这种比较技术，声音识别系统可以确定出存储的声音中的与新声音的最优匹配，并且从而可以将该新声音与一组 TTS 标记参数相关联。下面的描述描述这些和类似技术的实施例，以及所示网络 100 的实施例的组件可执行这些功能的方式。

在所示出的网络 100 的实施例中，网络 102 代表任何硬件和/或软件，用于在组件之间传输声音和/或数据信息，其中这些信息采用分组、帧、信元、段或数据的其他部分（通称为“分组”）的形式传输。网络 102 可包括路由器、交换机、集线器、网关、链路和其他合适的硬件和/或软件组件的任何组合。网络 102 可使用用于传输信息的任何合适的协议或介质，包括因特网协议（IP）、异步传输模式（ATM）、同步光网络（SONET）、以太网、或者任何其他合适的通信介质或协议。

网关 106 耦合网络 102 到 PSTN 104。一般来说，网关 106 代表任何这样的组件，其用于将适于网络 102 传输的一种格式的信息转变为适于在任何其他类型的网络中传输的另一种格式。例如，网关 106 可以将来自数据网络 102 的分组化的信息转换为在 PSTN 104 上传输的模拟信号。

端点 108 代表任何这样的硬件和/或软件，其用于接收来自用户的任何合适形式的信息，将这种信息传输到网络 100 的其他组件，并且将接收自网络 100 的其他组件的信息呈现给其用户。端点 108 可包括电话、IP 电

话、个人计算机、声音软件、显示器、麦克风、扬声器或任何其他合适形式的信息交换设备。在特定的实施例中，端点 108 可包括用于执行涉及信息传输的其他任务的处理能力和/或存储器。

SFV 服务器 200 代表包括硬件和/或软件的任何这样的组件，其分析语音信号，并且计算一系列时间段的语音的声学特征、一类声音特征向量。SFV 服务器 200 可接收任何合适形式的语音，包括模拟信号、来自麦克风的直接语音输入、分组化的语音信息，或者包括任何合适的用于将语音采样传输到 SFV 服务器 200 的方法。SFV 服务器 200 可使用任何合适的技术、方法或算法来分析所接收到的语音。

在特定实施例中，SFV 服务器 200 计算用于修正的高斯混合模型（GMM）的语音特征向量，例如在由 Douglas A. Reynolds、Thomas F. Quatieri 和 Robert B. Dunn 著的“Speaker Verification Using Adapted Gaussian Mixture Models”和由 Douglas A. Reynolds 和 Richard C. Rose 著的“Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”中所述的那些。在这种高斯混合模型分析的特定实施例中，通过确定具有递增带宽的对数空间滤波器（“mel-滤波器”）的频谱能量，从而计算出语音特征向量。从而获得的 log-频谱能量的离散余弦变换被称作语音的“mel-刻度式倒频谱”。mel-刻度式倒频谱中的项的系数被称作“特征向量”，它们被归一化来消除线性通道卷积效果（加性偏置，additive bias），并且计算这些特征向量的不确定性范围（“ δ 倒频谱”）。例如，通过倒频谱平均值消去法（CMS）和/或相对频谱（RASTA）处理，可以消除加性偏置。使用诸如在相邻特征向量的范围内拟合多项式之类的技术，可以计算出 δ 倒频谱。所产生的特征向量定义了声音的特征，并且可以使用各种统计分析技术来与其他声音相比较。

声音匹配服务器 300 代表任何合适的硬件和/或软件，其用于将所测得的参数集合与语者模型（speaker model）相比较，并且确定出所测得的语音特征向量和语者模型之间的优选匹配。“语者模型”指描述由文本到语音设备或算法所产生的声音的任何数学量或量的集合。语者模型可被挑选为与 SFV 服务器 200 所确定出的语音特征向量的类型一致，以便帮助实现

语者模型和测得的语音特征向量之间的比较，并且它们可响应于特定的文本消息、声音采样或其他源而被存储或产生。声音匹配服务器 300 可采用任何合适的技术、方法或算法，来比较所测得的语音特征向量和语者模型。例如，声音匹配服务器 300 可以使用相似性函数，例如高斯混合模型的 log-相似性函数或更复杂的隐马尔可夫模型的相似性函数来匹配语音特征。在特定实施例中，声音匹配服务器 300 使用高斯混合模型来比较测得的参数与声音模型。

也可以采用各种其他语音分析技术。例如，诸如音调频谱再现之类的声学特征长时间平均 (long-term averaging) 可以揭示出语音的独特特征，这是通过移除使得难以标识出说话者的语音变化和其他短时间语音效果实现的。其他技术包括基于类似文本来从语音上比较发声，以标识出声音的不同特征。这些技术可以使用隐马尔可夫模型 (HMM)，通过考虑音素之间的潜在关系（“马尔可夫连接”），从而分析类似的音素之间的差别。替换技术可以包括在神经网络中训练识别算法，以使所使用的识别算法可取决于该网络所针对训练的特定说话者而变化。网络 100 可适于使用任何描述的技术或任何合适的技术，以使用测得的语音特征向量来针对一组候选语者模型中的每个计算分值，并且确定出测得的语音特征向量与语者模型中的一个之间的最优匹配。“语者模型”指任何这样的数学量，这些数学量定义与 TTS 标记参数的特定集合相关联的声音的特征，并且被用在最优匹配的测得的语音向量的假设检验中。例如，对于高斯混合模型，语者模型可包括混合密度函数中的高斯数、N 概率权重的集合、每个成员高斯密度的 N 平均值向量的集合，以及每个成员高斯密度的 N 协方差矩阵的集合。

TTS 服务器 400 代表任何这样的硬件和/或软件，其用于从文本信息产生声音信息。可以产生任何合适输出形式的声音信息，包括模拟信号、自扬声器输出的声音、分组化的音频信息、或者任何其他用于传输声音信息的合适格式。由 TTS 服务器 400 创建的声音信息的声学特征利用 TTS 标记参数而被控制，这些参数可包括用于所提供的音频的各种声学属性的控制信息。文本信息可存储为任何合适的文件格式，包括电子邮件、即时消

息、存储的文本文件、或者任何其他信息的机器可读形式。

统一的消息发送服务器 110 代表包括硬件和/或软件的任何这样的网络中的一个或多个组件，其管理许多用户的不同类型的信息。例如，统一的消息发送服务器 110 可以维护网络 102 的用户的声音消息和文本消息。统一的消息发送服务器 110 还可以存储用户概况，包括提供对用户的声音的最接近匹配的 TTS 标记参数。统一的消息发送服务器 110 可由网络连接和/或声音连接访问，这允许用户登录或拨入到统一的消息发送服务器 110 来提取消息。在特定实施例中，统一的消息发送服务器 110 也可以维护用户的关联概况，这些关联概况包含关于这些用户的这样的信息，该信息有助于向网络 102 的用户提供消息发送服务。

在操作中，发送端点 108a 向接收端点 108b 发送文本消息。接收端点 108b 可以被设置为文本到语音模式，以使其将文本消息输出为语音。在该情形中，网络 100 的组件确定出与文本消息的源相关联的声音的语音特征向量集合。该文本消息的“源”可以指端点 108a，或者生成该消息的其他组件，并且也可以指这种设备的用户。因此，与文本消息的源相关联的声音例如可以是端点 108a 的用户的声音。网络 100 比较语音特征向量的集合与语者模型，来选择最优匹配，该最优匹配是指无论使用任何比较测试，该语者模型都被认为是该声音的语音特征向量集合的最优匹配。然后，网络 100 基于与被挑选为最优匹配的语者模型相关联的 TTS 标记参数来生成语音。

在一种操作模式中，网络 100 的组件检测到端点 108b 被设置为将文本消息作为声音消息接收。或者，当端点 108 被设置为将文本消息输出为声音消息时，端点 108b 可以将文本消息传输到 TTS 服务器 400。TTS 服务器 400 向发送该文本消息的端点 108b 发送请求声音采样的请求。SFV 服务器 200 接收到声音采样，并且分析该声音采样来确定该声音采样的语音特征向量。SFV 服务器 200 将语音特征向量传输到声音匹配服务器 300，该服务器然后将所测得的语音特征向量与声音匹配服务器 300 中的语者模型相比较。声音匹配服务器 300 确定出语者模型的最优匹配，并且通知 TTS 服务器 400，告知与优选的语者模型相关联的适当 TTS 标记参

数，以便 TTS 服务器 400 用来生成声音。然后，TTS 服务器 400 使用所选出的参数集合来生成此后自接收端点 108b 接收到的文本消息的声音。

在另一操作模式中，TTS 服务器 400 可以向发送端点 108a 请求一组定义声音特征的语音特征向量。如果这种兼容的语音特征向量是可获得的，则声音匹配服务器 300 可以直接从发送端点 108a 接收到这些语音特征向量，并且将那些语音特征向量与声音匹配服务器 300 存储的语者模型相比较。这样，声音匹配服务器 300 通过与发送端点 108a 交换信息来确定出与所采样的声音最佳匹配的语者模型设置。

在另一操作模式中，声音匹配服务器 300 可以使用 TTS 服务器 400 来生成语者模型，这些模型然后用在源的语音特征向量的假设检验中，这和 SFV 服务器 200 所确定的一样。例如，存储的声音采样可在发送端点 108a 被关联到具体的文本。在那种情形中，SFV 服务器 200 可接收到声音采样，并且对其进行分析，而声音匹配服务器 300 接收到文本消息。声音匹配服务器 300 将文本消息传输到 TTS 服务器 400，并且指示 TTS 服务器 400 根据可用 TTS 标记参数阵列基于该文本消息生成声音数据。每个 TTS 标记参数集合对应于声音匹配服务器 300 中的语者模型。这根据相同的文本片断有效地产生许多不同的声音。然后，SFV 服务器 200 分析各声音采样，并且计算声音采样的语音特征向量。SFV 服务器 200 将这些语音特征向量传输到声音匹配服务器 300，声音匹配服务器 300 使用这些语音特征向量对候选语者模型执行假设检验，这些模型中的每一个对应于特定 TTS 标记参数集合。由于这些声音采样是从相同的文本生成的，所以在从端点 108a 接收到的声音与模型声音相比时，可以实现更高的准确度。

所述的用于确定对应于实际声音的准确模型的操作模式和技术可以实现在多种不同的实施例中。在替换实施例的一种示例中，在分布式通信体系统结构中的端点 108 包括足以执行所述服务器 200、300 和 400 的任何或全部任务的功能。因此，设置为将文本信息输出为声音信息的端点 108 可执行下述步骤：获取声音采样、确定用于 TTS 生成的匹配 TTS 标记参数集合、以及使用所选出的参数集合产生语音输出。在这种实施例中，端点 108 也可以分析它们各自的声音，并且维护可被传输到兼容的声音

识别系统的语音特征向量集合。

在另一替换实施例中，所述技术可用在统一的消息发送系统中。在这种情形中，服务器 200、300 和 400 可与统一的消息发送服务器 110 交换信息。例如，统一的消息发送服务器 110 可以维护作为特定用户的概况一部分的声音采样。在此情形中，SFV 服务器 200 和声音匹配服务器 300 可以使用存储的每个用户的采样和/或参数来确定该用户的准确匹配。这些操作可在网络 102 中本地执行，或者与使用统一的消息发送服务器 110 的远程网络协作执行。这样，这些技术可适于广泛的消息发送系统。

在其他替换实施例中，SFV 服务器 200、声音匹配服务器 300 和 TTS 服务器 400 的功能可被集成或分布在多个组件中。例如，网络 102 可包括执行所述声音分析和模型选择任务中的任何任务和全部任务的混合服务器。在另一示例中，TTS 服务器 400 可以代表这样的独立服务器的集合，这些服务器中的每个都根据特定的 TTS 标记参数集合生成语音。因此，声音匹配服务器 300 可以选择与所选出的 TTS 标记参数集合相关联的特定服务器 400，而不是将特定的参数集合传输到 TTS 服务器 400。

本发明的某些实施例的一个技术优点在于对于端点 108 的用户的增加的用途。使用与提供文本消息的人的声音类似的声音为特定端点 108 的用户提供了增加的能力，使其能够识别出使用次要队列的源。一般来说，该特征通常也可以使用户更容易地与网络 100 中的 TTS 系统交互。

某些实施例的另一技术优点在于与其他系统的互操作能力。由于端点 108 已配备为交换声音信息，所以端点 108 不需要额外的硬件、软件或共享协议来向 SFV 服务器 200 或声音匹配服务器 300 提供声音采样。因此，所述技术可以被吸收到现有系统中，结合不使用相同的语音分析和再现技术的系统一起工作。

图 2 示出了 SFV 服务器 200 的特定实施例。在所示实施例中，SFV 服务器 200 包括处理器 202、存储器 204、网络接口 206 和语音接口 208。一般来说，SFV 服务器 200 对 SFV 服务器 200 接收到的声音执行分析，并且产生描述所接收到的声音的音频特征的数学量（特征向量）。

处理器 202 代表用于处理信息的任何硬件和/或软件。处理器 202 可包

括微处理器、微控制器、数字信号处理器（DSP）、或者任何其他合适的硬件和/或软件组件。处理器 202 执行存储在存储器 204 中的代码 210 来执行 SFV 服务器 200 的各种任务。

存储器 204 代表任何形式的信息存储装置，无论是易失性的还是非易失性的。存储器 204 可包括光介质、磁介质、本地介质、远程介质、可移除介质、或者任何其他合适的信息存储形式。存储器 204 存储由处理器 202 执行的代码 210。在所述示例中，代码 210 包括特征确定算法 212。算法 212 代表用于数学地定义声音信息的特征的任何合适的技术或方法。在特定实施例中，特征确定算法 212 对语音进行分析，并且计算在用于语音比较的高斯混合模型中使用的一组特征向量。

接口 206 和 208 代表任何端口或连接，不管是真正的还是虚拟的，它们允许 SFV 服务器 200 与网络 100 的其他组件交换信息。网络接口 206 用来与数据网络 102 的组件交换信息，这些组件包括在上述操作模式中描述的声音匹配服务器 300 和/或 TTS 服务器 400。语音接口 208 允许 SFV 服务器 200 接收语音，不管是通过麦克风，还是以模拟形式、分组形式或者任何其他合适的声音传输方法。语音接口 208 可以允许 SFV 服务器 200 与端点 108、统一的消息发送服务器 110、TTS 服务器 400 或可使用 SFV 服务器 200 的语音分析能力的任何其他组件交换信息。

在操作中，SFV 服务器 200 在语音接口 208 处接收到语音数据。处理器 202 执行特征确定算法 212 来确定出定义语音特征的语音特征向量。SFV 服务器 200 使用网络接口 206 将语音特征向量传输到网络 100 的其他组件。

图 3 示出了声音匹配服务器 300 的一个实施例的示例。在所示实施例中，声音匹配服务器 300 包括处理器 302、存储器 304 和网络接口 306，它们与上述 SFV 服务器 200 的类似组件相似，并且可包括结合图 2 中的类似组件所描述的任何硬件和/或软件。声音匹配服务器 300 的存储器 304 存储代码 308、语者模型 312 和接收到的语音特征向量 314。

代码 308 代表这样的指令，处理器 302 执行这些指令来执行声音匹配服务器 300 的任务。代码 308 包括比较算法 310。处理器 302 使用比较算

法 310 来将一组语音特征向量与语者模型的集合相比较，以确定所考虑的语音特征向量集合与这些模型之一之间的优选匹配。比较算法 310 可以是假设检验算法，在该算法中，给予所建议的匹配一个匹配所考虑的语音特征向量集合的概率，但是也可以包括任何其他合适类型的比较。语者模型 312 可以是基于先前利用 TTS 服务器 400 生成的可用声音进行的训练的已知参数集的集合。或者，语者模型 312 可以是按照来自源端点 108 的要被转换为语音的特定文本消息的需求而基于每种情形所生成的。接收到的语音特征向量 314 代表这样的参数，这些参数定义与来自其的文本要被转换为语音的源端点 108 相关联的声音采样的特征。接收到的语音特征向量 314 一般是上述 SFV 服务器 200 执行的分析的结果。

在操作中，声音匹配服务器 300 使用网络接口 306 从 SFV 服务器 200 接收到语音特征向量，这些语音特征向量定义与端点 108 相关联的声音的特征。处理器 302 在存储器 304 中存储参数，并且执行比较算法 310 来确定所接收到的语音特征向量 314 与语者模型 312 之间的优选匹配。处理器 302 从语者模型 312 中确定出优选匹配，并且将关联的 TTS 标记参数传输到 TTS 服务器 400，这些参数将要用于随后从接收自特定端点 108 的文本消息生成语音中。也可以使用替换操作模式。例如，声音匹配服务器 300 可以在从 SFV 服务器 200 接收到接收到的语音特征向量 314 之后生成语者模型 312，而不是维护存储的语者模型 312。这可以在确定语者模型 312 中的优选匹配时提供额外的通用性和/或准确性。

图 4 示出了 TTS 服务器 400 的特定实施例。在所示出的实施例中，TTS 服务器 400 包括处理器 402、存储器 404、网络接口 406 和语音接口 408，它们与结合图 2 所述的 SFV 服务器 200 的类似组件相似，并且可包括其中所述的任何硬件和/或软件。一般地说，TTS 服务器 400 接收文本信息，并且使用 TTS 引擎 412 从该文本生成声音信息。

TTS 服务器 400 的存储器 404 存储代码 410 和存储的 TTS 标记参数 414。代码 410 代表由处理器 402 执行来执行 TTS 服务器 400 的各种任务的指令。代码 410 包括 TTS 引擎 412，其代表从声音数据产生语音的技术、方法或算法。所使用的特定 TTS 引擎 412 可取决于声音信息的可用输

入格式和期望输出格式。TTS 引擎 412 可适用于多种文本格式和声音输出格式。TTS 标记参数 414 代表 TTS 引擎 412 用来生成语音的参数集合。取决于所选出的 TTS 标记参数 414 的集合，TTS 引擎 412 可以产生具有不同发声特性的声音。

在操作中，TTS 服务器 400 基于使用网络接口 406 接收到的文本消息生成语音。使用语音接口 408，该语音被传输到端点 108 或其他目的地。为了生成特定文本消息的语音，向 TTS 服务器 400 提供特定的 TTS 标记参数 414 集合，并且相应地使用 TTS 引擎 412 生成语音。在 TTS 服务器 400 未将特定声音关联到消息的情形中，TTS 服务器 400 可以使用与默认声音相对应的 TTS 标记参数 414 的默认集合。当依赖于源的信息可用时，TTS 服务器 400 可以从声音匹配服务器 300 接收到适当的 TTS 标记参数选择，以使 TTS 标记参数对应于优选语者模型。这可以允许 TTS 服务器 400 产生对发送文本消息的人的声音的更准确的再现。

图 5 示出了端点 108b 的特定实施例。在所示出的实施例中，端点 108b 包括处理器 502、存储器 504、网络接口 506 和用户接口 508。处理器 502、存储器 504 和网络接口 506 对应于前述 SFV 服务器 200、声音匹配服务器 300 和文本到语音服务器 400 的相似组件，并且可包括与前述那些组件的硬件和/或软件相似的任何硬件和/或软件。用户接口 508 代表任何这样的硬件和/或软件，端点 108b 利用这些硬件和/或软件与用户交换信息。例如，用户接口 508 可包括麦克风、键盘、小键盘、显示器、扬声器、鼠标、图形用户界面、按钮或者信息交换的任何其他合适形式。

端点 108b 的存储器 504 存储代码 512、语者模型 518、以及接收到的语音特征向量 520。代码 512 代表由处理器 502 执行来执行端点 108b 的各种任务的指令。在特定实施例中，代码 512 包括特征确定算法 512、比较算法 514 和 TTS 引擎 516。算法 512 和 514 以及引擎 516 分别对应于结合 SFV 服务器 200、声音匹配服务器 300 和 TTS 服务器 400 所述的类似算法。因此，端点 108b 将那些组件的功能集成到了单个设备中。

在操作中，端点 108 使用网络接口 506 与网络 100 的其他端点 108 和/或组件交换声音和/或文本信息。在与其他设备交换声音信息期间，端点

108b 可以使用特征确定算法 512 确定出接收到的语音的语音特征向量 520，并且在存储器 504 中存储那些特征向量 520，从而将参数 520 关联到发送端点 108a。端点 108b 的用户可以触发端点 108b 的文本到语音模式。在文本到语音模式中，端点 108b 使用 TTS 引擎 516 从接收到的文本消息生成语音。端点 108b 通过使用比较算法 514 来将参数 520 与语者模型 518 相比较，从而选择出用于基于文本消息的源生成语音的语者模型集合 518，并且使用与优选模型相关联的 TTS 标记参数来生成语音。这样，TTS 引擎 516 所产生的语音紧密对应于文本消息的源。

在替换实施例中，端点 108b 可以执行不同的或额外的功能。例如，端点 108b 可使用特征确定算法 512 来分析其自己的用户的语音。该信息可与其他端点 108 交换并且/或者与语者模型 518 相比较来提供依赖于源的文本到语音的协作方法。类似地，端点 108 可以协作地协商出一组语者模型 518，以用在文本到语音操作中，这允许分布式网络体系结构确定合适的协议来允许依赖于源的文本到语音处理。一般来说，端点 108 的描述可以与前面任何地方描述的网络 100 的任何实施例一致。

图 6 示出了一种这样的方法的流程图 600，该方法选择一组适当的 TTS 标记参数，以在网络 100 中产生依赖于源的语音输出。在步骤 602 中，端点 108 接收到文本消息。如果端点 108 具有能够将文本转换为声音的设置，则消息可由端点 108 接收到，并且被传输到网络 100 中的其他组件，或者，可以被 TTS 引擎 400 或另一个组件接收到。在判定步骤 604 中，确定出端点 108 是否具有所选择的 TTS 选项。如果端点 108 不具有所选择的 TTS 选项，则在步骤 606 中，消息以文本形式被传输到端点。如果已选择了 TTS 选项，则在步骤 608 中，TTS 引擎 400 确定出是否可获得语音特征向量。这可以是先前已针对发送消息的端点 108 确定了语音特征向量的情形，或者是在端点 108 使用兼容声音特征系统时维护端点 108 的用户的语音特征向量的情形。如果语音特征向量是不可获得的，TTS 引擎 400 接下来在判定步骤 610 中确定是否可获得语音采样。如果语音特征向量和语音采样都不可获得，则在步骤 612 中 TTS 引擎 400 使用默认 TTS 标记参数来定义语音的特征。

如果语音采样可获得，则 SFV 服务器 200 在步骤 614 中分析该语音采样来确定该声音采样的语音特征向量。在从端点 108 接收到特征向量或者由 SFV 服务器 200 确定出特征向量之后，声音匹配服务器 300 在步骤 616 中比较特征向量和语者模型，并且在步骤 618 中从那些参数确定出优选匹配。

在选择出语音特征向量的优选匹配或者使用默认的 TTS 标记参数集合之后，TTS 引擎 400 在步骤 620 中使用关联的 TTS 标记参数生成语音。TTS 引擎 400 在步骤 622 中使用语音接口 408 输出语音。然后，TTS 引擎 400 在判定步骤 624 中确定是否有额外的文本消息要被转换。作为步骤 624 的一部分，TTS 引擎 400 可以验证端点 108 是否仍被设置为以声音形式输出文本消息。如果存在来自端点 108 的额外的文本消息（或者如果端点 108 不再设置为以声音形式输出文本消息），则 TTS 引擎 400 使用先前选择出的参数来从后继文本消息生成语音。否则，该方法结束。

尽管已用多个实施例描述了本发明，但是可以向本领域的技术人员建议多种改变、变化、变更、变换和修改，并且本发明是要包括这些改变、变化、变更、变换和修改，只要它们落在所附权利要求的范围内。

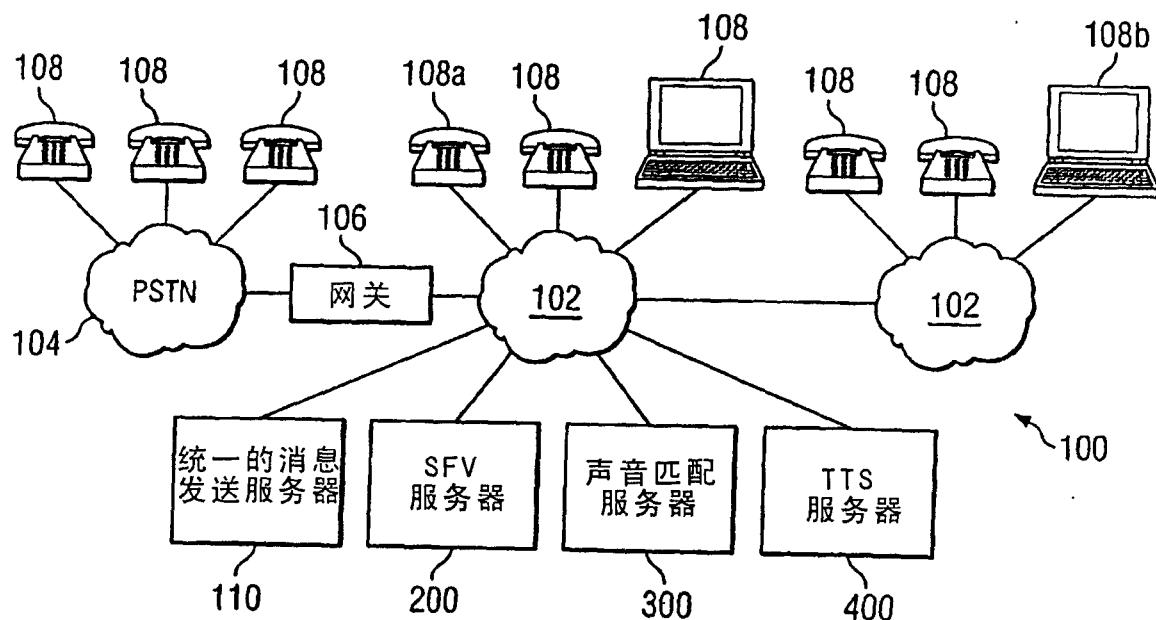


图1

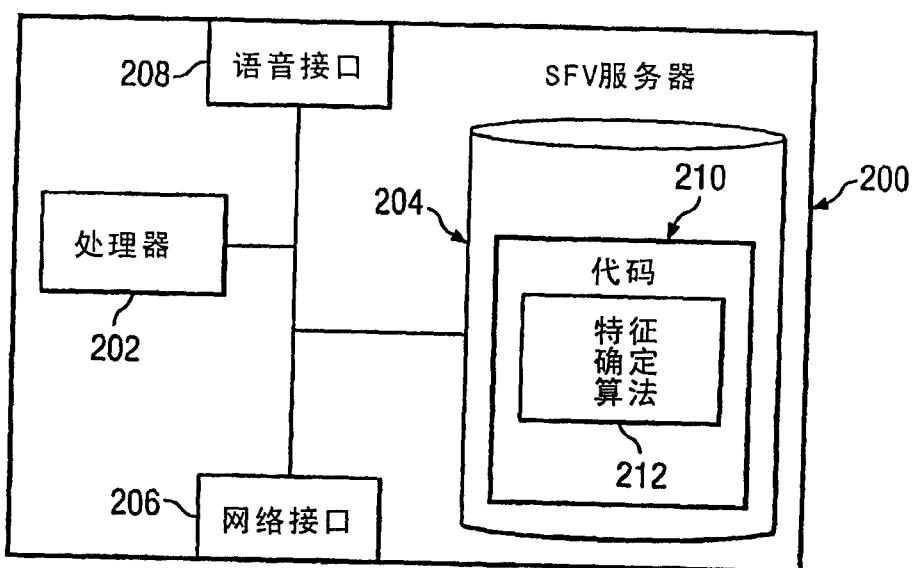


图2

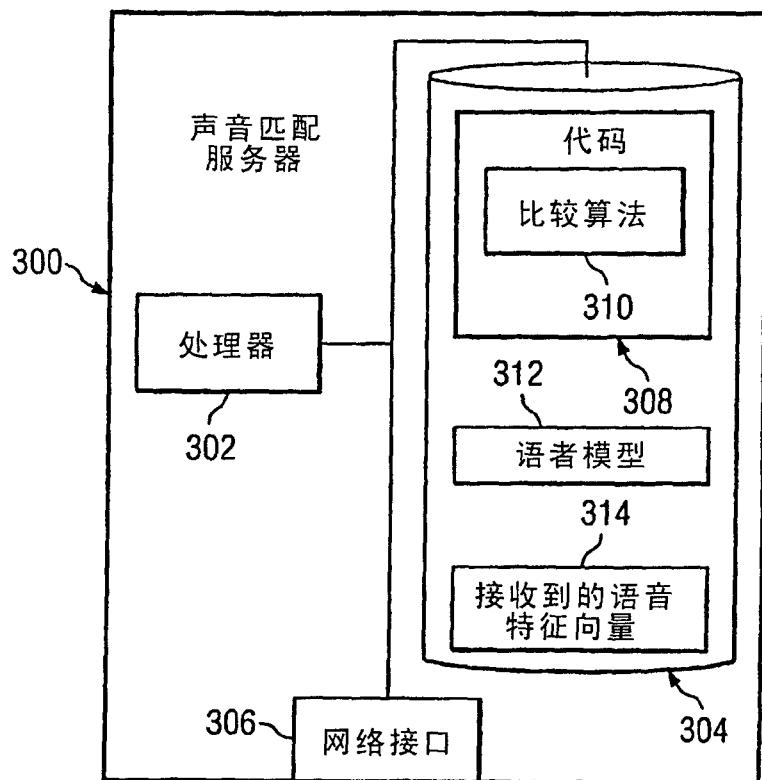


图3

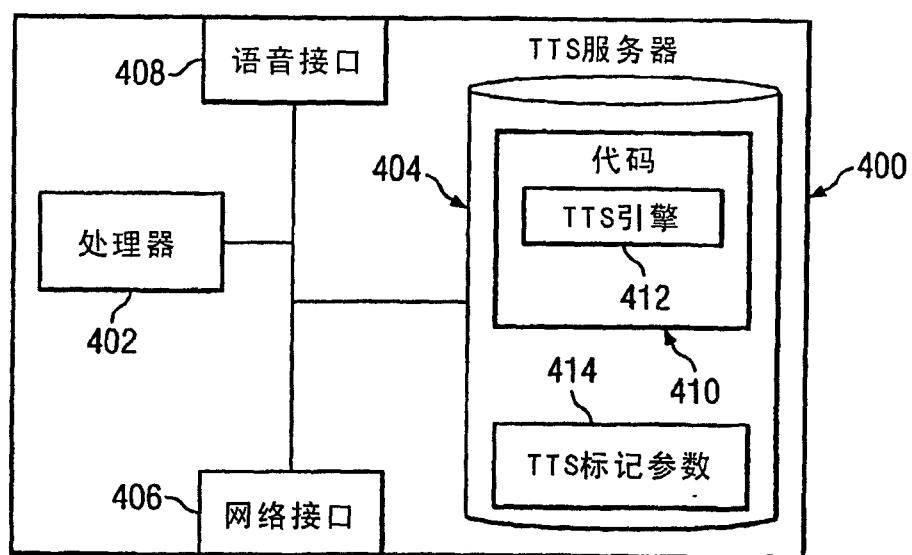


图4

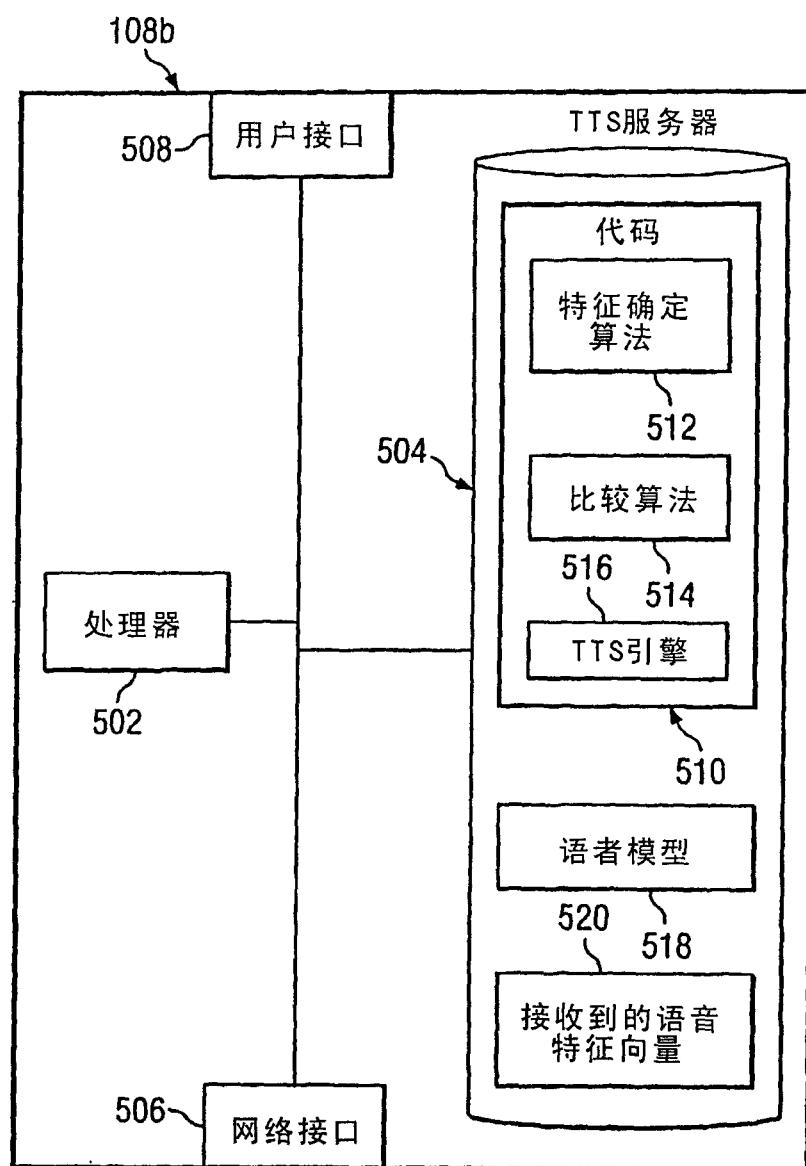


图5

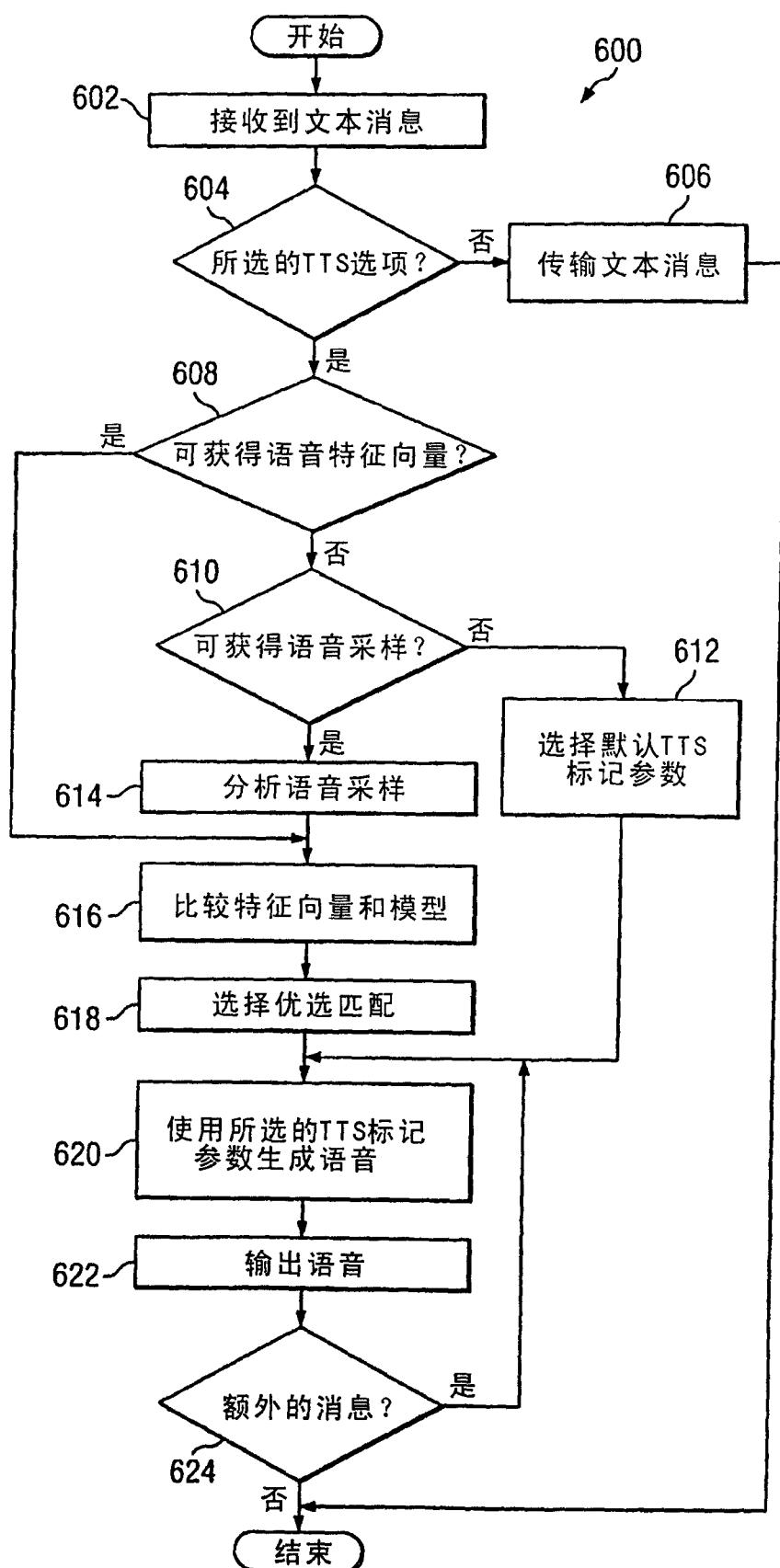


图6