(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 103650002 B
(45)授权公告日 2018.02.23

(21)申请号 201280033415.8
(22)申请日 2012.05.04
(65)同一申请的已公布的文献号
申请公布号 CN 103650002 A
(43)申请公布日 2014.03.19
(30)优先权数据
61/483,571 2011.05.06 US
(85)PCT国际申请进入国家阶段日
2014.01.06
(86)PCT国际申请的申请数据
PCT/US2012/036679 2012.05.04
(87)PCT国际申请的公布数据
WO2012/154618 EN 2012.11.15
(73)专利权人 西尔股份有限公司
地址 美国加利福尼亚州
(72)发明人 柏鲁兹·瑞斯维尼 阿里·卢西
(74)专利代理机构 北京商专永信知识产权代理事务所(普通合伙) 11400
代理人 邬玥 葛强
(51)Int.Cl.
G06T 13/40(2011.01)
H04M 1/725(2006.01)
G10L 13/10(2013.01)
(56)对比文件
US 2010/0302254 A1,2010.12.02,
US 7664645 B2,2010.02.16,
US 7298256 B2,2007.11.20,
CN 1117344 C,2003.08.06,
US 2010/0302254 A1,2010.12.02,
US 6970820 B2,2005.11.29,
US 6250928 B1,2001.06.26,
审查员 朱琳玲

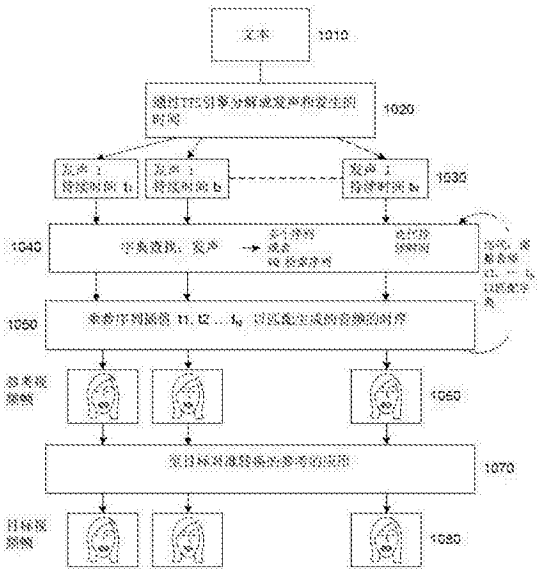
权利要求书3页 说明书13页 附图13页

(54)发明名称

基于文本的视频生成

(57)摘要

本文公开了用于生成基于文本序列的人的视频序列的技术。基于接收到的文本序列,处理装置生成人的视频序列以模拟人的视觉和听觉情感表达,包括使用人的声音的音频模型生成视频序列的音频部分。基于对人的先验知识,处理装置可以模拟在视频序列的视觉部分中的情感表达。例如,先验知识可以包括在现实生活中的人的照片或视频。



CN 103650002 B

1.一种生成视频序列的方法,包括:

在处理装置输入文本序列;以及

由所述处理装置基于所述文本序列生成真人的包括多个帧的视频序列以模拟所述真人的视觉和听觉情感表达,包括使用所述真人的声音的音频模型生成所述视频序列的音频部分,所述生成真人的视频序列是基于视觉模型,所述视觉模型采用所述真人的现实生活中的视频序列作为训练数据;

其中所述视频序列的每个帧由在与多个向量对应的乘数加权的真人的特征空间中的多个向量的线性组合来表示,并且每个乘数表示沿着特征空间中的相应向量的真人的均值图像的偏离程度。

2.如权利要求1所述的方法,其中所述处理装置是移动装置,所述文本序列从第二移动装置通过短消息服务信道被输入,以及所述生成真人的视频序列包括由所述移动装置基于存储在所述移动装置和所述第二移动装置上的共享信息生成真人的视频序列。

3.如权利要求1所述的方法,其中所述文本序列包括一组字,所述一组字包括至少一个字,以及其中所述视频序列被生成,以使得所述真人呈现在所述视频序列中说出所述字。

4.如权利要求1所述的方法,其中所述文本序列包括表示发声的文本,以及其中所述视频序列被生成产生,以使得所述真人呈现在所述视频序列中说出发声。

5.如权利要求1所述的方法,其中所述文本序列包含字和所述字的指标,当所述真人呈现在所述视频序列中说出所述字时,该指标指示在所述视频序列中的一个时间上的所述真人的情感表达,所述指标是在预定的指标集合之内,并且所述预定的指标集合中的每个指标是与不同的情感表达相关联。

6.如权利要求1所述的方法,其中所述生成视频序列包括:

由所述处理装置生成真人的视频序列以基于所述文本序列和所述真人的先验知识模拟所述真人的视觉和听觉情感表达。

7.如权利要求1所述的方法,其中所述生成视频序列包括:

映射在所述文本序列中的字至所述真人的面部特征;以及

在背景场景中呈现所述真人的所述面部特征。

8.如权利要求7所述的方法,其中所述字基于所述字的一个或更多个指标被映射到面部特征,其中所当所述真人呈现在所述视频序列中说出所述字时,所述指标指示在所述视频序列中的一个时间上的所述真人的情感表达。

9.如权利要求7所述的方法,其中所述面部特征包括专门适用于所述真人的特定面部特征。

10.如权利要求7所述的方法,其中所述生成所述视频序列还包括:生成与所述真人的面部特征相适应的所述真人的身体姿势。

11.如权利要求1所述的方法,其中所述生成视频序列包括:

通过使用基于所述真人的声音的所述音频模型,基于所述文本序列中的字生成表示所述真人的说话的音频序列。

12.如权利要求1所述的方法,其中所述文本序列的接收包括:

实时接收文本序列;

其中所述视频序列的生成包括:

基于所述文本序列实时生成真人的视频序列以模拟所述真人的视觉和听觉情感表达，包括使用所述真人的声音的音频模型来生成所述视频序列的音频部分。

13.如权利要求1所述的方法，其中所述真人的模拟的视觉和听觉情感表达独立于所述视觉模型的所述训练数据。

14.一种生成视觉序列的方法，包括：

在处理装置输入文本序列；

由所述处理装置基于所述文本序列生成真人的包括多个帧的视觉序列以模拟所述真人的视觉情感表达，其中所述生成真人的视频序列是基于视觉模型，所述视觉模型采用所述真人的现实生活中的视频序列作为训练数据，其中所述视觉序列的每个帧的面部是由与多个向量对应的乘数加权的真人的特征空间中的多个向量的线性组合来表示，并且每个乘数表示沿着特征空间中的相应向量的真人的均值图像的偏离程度，以及其中所述真人的特征空间具有与所述真人的面部特征的点相关联的尺寸；

利用所述真人的声音的声音模型，由所述处理装置基于文本序列生成所述真人的音频序列以模拟所述真人的发声情感表达，以及

由所述处理装置基于所述文本序列通过合并所述视觉序列和所述音频序列制造所述真人的视频序列，其中基于所述文本序列，所述视觉序列和音频序列是同步的。

15.如权利要求14所述的方法，其中所述视觉序列的每个帧的所述面部是由所述真人的先验图象的线性组合所表示，以及所述真人的所述先验图像的每一个先验图像对应于所述真人的均值图像偏差。

16.如权利要求14所述的方法，其中所述生成基于所述文本序列的真人的视觉序列包括：

将所述视觉序列的每个帧划分成两个或多个区域，其中至少一个的所述区域是由所述真人的先验图像的组合来表示。

17.如权利要求14所述的方法，其中所述真人的声音的所述音频模型包括多个从所述真人的语音样本中创建的声音特征，所述多个声音特征中的每个声音特征对应于一文本。

18.如权利要求17所述的方法，其中所述多个声音特征中的每个声音特征对应于字，音素，或发声。

19.如权利要求14所述的方法，其中所述真人的声音的所述音频模型包括多个声音特征，所述声音特征从所述真人的语音样本、第二真人的根据所述文本序列的语音，以及在所述真人的声音的波形与所述第二真人的声音波形之间的对应关系被创建；以及其中基于所述真人的声音波形和所述第二真人的声音波形之间的对应关系所述真人的声音特征被映射到所述第二真人的语音。

20.如权利要求14所述的方法，其中所述真人的声音的所述音频模型包括多个声音特征，所述声音特征从所述真人的语音样本，根据所述文本序列由文本—语音模型所产生的语音，以及所述真人的声音波形和所述文本—语音模型的声音波形之间的对应关系创建，以及其中基于所述真人的声音波形和所述文本—语音模型的声音波形之间的对应关系所述真人的声音特征被映射到所述语音。

21.一种处理用于视频生成的文本序列的方法，包括：

创建文本序列，其中所述文本序列代表一个或多个字，一个真人将要在包括多个帧的

视频序列中说出所述一个或多个字，所述视频序列使用基于所述真人的声音的音频模型被生成，所述视频序列视觉地和听觉地表示所述真人的情感表达范围，所述生成真人的视频序列是基于视觉模型，所述视觉模型采用所述真人的现实生活中的视频序列作为训练数据，其中视频序列的每个帧由在与多个向量对应的乘数加权的真人的特征空间中的多个向量的线性组合来表示，并且每个乘数表示沿着特征空间中的相应向量的真人的均值图像的偏离程度；

确定识别指标，其中所述指标为与所述文本序列中的字相关联的指标，所述指标是在预定指标集合中的一个，其中的每个预定指标指示所述真人的不同的情感表达；

将所述指标并入到所述文本序列；以及

发送所述文本序列至被配置为生成所述视频序列的装置。

22.如权利要求21所述的方法，其中所述确定识别指标包括：

从条目菜单中选择一个条目与所述文本序列中的字相关联，其中在所述菜单中的每个条目是提示所述真人的情感表达的指标。

23.如权利要求21所述的方法，其中所述确定识别指标包括：

使用自动语音识别引擎基于讲话者读出所述文本序列中的所述字的音频序列识别与所述文本序列中的字相关联的指标。

基于文本的视频生成

[0001] 对相关申请的交互引用

[0002] 本申请要求2011年05月06日递交的第61/483,571号美国临时申请的优先权，该临时申请以引用方式被合并于此。

技术领域

[0003] 本发明的至少一个实施例涉及视频生成，并且更具体地涉及用于生成视频以基于文本序列(如用户-用户的短文本消息)真实地模拟人的视觉和听觉情感表达的方法和系统。

背景技术

[0004] 数字视频序列可包含非常大的数据量。要利用现有技术有效地传输视频，必须需要大的传输带宽。然而，通过无线介质的数据传输带宽是有限的且往往是昂贵的资源。

[0005] 例如，短消息服务(“SMS”)，有时也被称为“发短信”，是目前使用最流行的人对人的通信技术之一。短信功能被广泛使用在几乎所有的现代手机。但是，SMS具有非常有限的能力来传递信息，每个短信有140字节或160个字符的固定长度，因此不适合于传输视频数据。多媒体讯息服务(“MMS”)是发送包含多媒体内容的消息的可能方式。然而，MMS消息不能利用现有的SMS基础设施，所以它花费比短信更高的成本。在非常低的带宽信道(如SMS通道)上视频消息即便并非完全不可能也是非常困难的。

发明内容

[0006] 在此公开的技术包括基于文本序列生成人的视频序列的技术和装置以真实地模拟根据文本序列的人的讲话，包括基于文本序列真实地模拟人的视觉和听觉情感表达。该技术能够产生已经通过低带宽信道(例如，SMS信道)发送的人的视频而实际上不需要通过该信道传输的视频数据的外观。

[0007] 此外，本技术提供了一种具有成本效益的方法来生成人讲话(或唱歌，或进行任何其它类型的发声)的现实的、非动画的视频。当人因各种原因，包括日程安排的冲突，不愿意，死亡和紧急情况不可用时，该技术提供了一种解决方案以替代捕捉人的视频。除了有关被存储在处理装置中的人的先验信息，该技术只需要一个文本序列，以便能够产生视频以及如果文本序列需要被传输仅需要很少的数据传输。文本序列提供了一种机制，通过调整文本序列的内容来控制和调整字和人似乎在视频序列中说出的情感表达。

[0008] 在一个实施例中，处理装置接收文本序列。基于接收到的文本序列，处理装置生成人的视频序列，以模拟人的视觉和听觉情感表达，包括使用人的声音的音频模型生成视频序列的音频部分。在视频序列的视觉部分的情感表达是基于有关人的先验知识被模拟。例如，先验知识可以包括在实际生活中捕获的人的照片和/或视频。

[0009] 在某些实施例中，文本序列包括一组字，该组字包括至少一个字。视频序列被产生以使得人似乎在视频序列中说出的字。文本序列还可以包括与一个或多个字相关联的情感

指标。每个指标指示的人在视频序列中的一个时间上的情感表达，当人似乎说出在视频序列中的指标相关联的字。在某些实施例中，处理装置将文本序列中的字映射到的人的面部特征，基于关于人的先验知识。稍后处理装置在背景场景中转换人的面部特征。

[0010] 在此介绍的技术其它方面通过附图以及随后的详细说明将是显而易见的。

附图说明

[0011] 本发明的这些和其它目的、特征和特性对于本领域的技术人员而言通过以下与所附权利要求和附图(所有这些都构成说明书的一部分)相结合的详细描述的研究将变得更加明显。在附图中：

[0012] 图1A示出了从发送装置接收文本序列的处理装置的例子；

[0013] 图1B示出了从发送装置经由中间装置接收文本序列的处理装置的例子；

[0014] 图2示出了从处理装置的输入组件接收文本序列的的处理装置的例子；

[0015] 图3是显示了处理装置的体系结构例子的高层次框图；

[0016] 图4示出了文本到视频(TTV)系统体系结构的示例；

[0017] 图5示出了构建人的视觉模型的例子；

[0018] 图6示出了将目标人物的面部划分成两个区域的例子的过程；

[0019] 图7示出了字典条目的例子；

[0020] 图8示出了创建文本到讲话(TTS)音频模式的例子的过程；

[0021] 图9示出了TTS音频合成的示例过程；

[0022] 图10示出了在视觉和听觉序列之间的同步的示例过程；

[0023] 图11示出了通过减少边界拟合误差将合成的视频嵌入背景的示例性过程；以及

[0024] 图12示出了基于2段模型将合成的视频嵌入到背景的示例性过程。

具体实施方式

[0025] 在本说明书中提及的“实施例”、“一个实施例”，等等，意味着被描述的特定特征、结构或特性被包括在本发明的至少一个实施例中。在本说明书中这样的短语的出现不一定都指代相同的实施例。

[0026] 一些相关文献，包括：K.皮尔逊的“On Lines and Planes of Closest Fit to Systems ofPoints in Space”，哲学杂志，1901年，第6卷系列2：第559-572页；T.F.库茨和C.J.泰勒的“Statistical Models of Appearance for Computer Vision”，曼彻斯特大学技术报告，2004年，第125页；T.库茨、G.爱德华和C.泰勒的“Active appearance models”，计算机视觉欧洲会议论文集，第2版：第484-489页，1998年；V.布兰兹和T.维特的“A morphable model for the synthesis of3d faces”，计算机图形和交互技术第26届年会论文集，187-194页，ACM出版/Addison-Wesley出版社出版有限公司，1999年；埃里卡.常和克里斯.布雷格勒的“Head Emotion”，斯坦福大学计算机科学技术报告，CSTR2003-02；F.J.黄和T.陈的“Real-time lip-synch face animation driven by human voice”，多媒体信号处理研IEEE讨会，1998年；马修斯.I、库茨.T.、班厄姆.A.、考克斯.S.和哈维.R.的“Extraction of visual features for lipreading”，模式分析与机器智能的IEEE交易，2002年，24(2)：第198-213页；N.李、S.德特默和M.沙阿的“Lipreading using

eigensequences”，自动脸部识别手势研讨会内部论文集，1995年，第30-34页；克里斯托夫.布雷格勒、米歇尔.科维尔和马尔科姆.斯拉尼的“Driving Visual Speech with Audio”，国际图形学年会论文集97版，第353-360页，1997年8月；所有文献的全部内容都通过引用方式被合并于此。

[0027] 图1A示出了处理装置以及在此所介绍的技术可以实施于其中的环境的例子。在图1A中，处理装置100经由互连120连接到发送装置110。互连120可以是，例如，蜂窝电话网络、SMS信道、电视频道、局域网(LAN)、广域网(WAN)、城域网(MAN)、全局区域网络(例如因特网)、光纤通道架构，或这些互连的任意组合。发送装置110能够经由互连120发送文本序列140到处理装置。处理装置100接收文本序列140，并基于文本的序列140生成视频序列150。任何发送装置110和处理装置100可以是，例如，蜂窝电话、传统的个人计算机(PC)、服务器级计算机、工作站，手持计算/通信装置，游戏控制台，电视等。

[0028] 处理装置110可以包括存储装置160用于存储所生成的视频序列150。存储装置160可以是，例如，传统的动态随机存取存储器(DRAM)，传统磁盘或光盘或磁带驱动器，非易失性固态存储器，诸如闪速存储器，或者这些装置的任意组合。

[0029] 任何的处理装置100和发送装置110可以包含操作系统(101，111)，用于管理处理装置100和发送装置110的操作。在某些实施例中，操作系统101和111是以软件形式实现的。然而，在其它实施例中，任何一个或多个的这些操作系统101和111可以以纯硬件被实现，例如，特别设计的专用电路或部分地以软件和部分作为专用电路来实现。

[0030] 文本序列诸如图1中的文本序列140可以包括指标(也叫标签，情感指标，或情感标签)。每一个指标指示当人似乎说出在视频序列中的这个字时，在视频序列中的人在一个时间上的情感表达。该指标可以是不同的形式，并且可以通过不同的方式来选择。在一个实施例中，指标从与文本序列内的字相关联的条目菜单中被选择作为一个条目，其中，在菜单中的每一个条目是表明人的情感表达的指标。在另一实施例中，指标是通过插入与文本序列内的字相关联的标记语言字符串被识别的，其中标记语言字符串来自标记语言字符串预定集合，以及在标记语言字符串的预定集合中的每一个标记语言字符串是表明人的情感表达的指标。仍在另一实施例中，指标在讲话者的音频序列上被识别，该讲话者通过使用自动语音识别(ASR)引擎读出在文本序列中的该字。

[0031] 图1B示出了从发送装置经由中间装置接收文本序列的处理装置的例子。在图1B中，处理装置100经由互连192连接到中间装置180。发送装置110经由互连191连接到中间装置180。任何互连191和192可以是，例如，蜂窝电话网络，SMS信道，电视频道，局域网(LAN)，广域网(WAN)，城域网(MAN)，全球地区网络(如因特网)，光纤通道架构，或这些互连的任意组合。在一些实施例中，互连191和192可以在一个网络中，例如，因特网。发送装置110能够经由互连191发送文本序列140的中间装置180的。中间装置经由互连192进一步发送该文本序列140到处理装置100。处理装置100接收到的文本序列140，并生成基于文本的序列140的视频序列150。中间装置可以是，例如，蜂窝电话、传统的个人计算机(PC)、服务器级计算机、工作站、手持计算/通信装置、游戏控制台、电视，等等。

[0032] 在一些实施例中，中间服务器180接收文本序列并处理该文本序列140为一组数据。该组数据而不是文本序列140被发送到处理装置100。

[0033] 图2示出了从处理装置的输入组件接收文本序列的处理装置的另一个例子。处理

装置200包括输入组件210，该输入组件210能够从人290接收文本序列240。处理装置200可以是，例如，蜂窝电话、传统的个人计算机(PC)、服务器级计算机、工作站、手持计算/通信装置、游戏控制台、电视，等等。输入组件210可以是，例如，键盘、鼠标、图像或视频照相机、麦克风、游戏控制台的控制器、遥控器、传感器、扫描仪、乐器或者这些装置的任意组合。

[0034] 该处理装置还包括基于文本序列240与人270的先验知识生成人的视频序列250的处理器205。视频序列250模拟了人的视觉和听觉情感表达，并且人似乎说出在视频序列250中的文本序列240内的特定字。生成的视频序列250可被存储在处理装置200内的存储装置260中。存储装置260可以是，例如，传统的动态随机存取存储器(DRAM)、传统的磁盘或光盘或磁带驱动器、非易失性固态存储器，诸如闪速存储器，或者这些装置的任意组合。文本序列240和/或人270的先验知识也可以被存储在存储装置260中，或者存储在与存储装置260分离的其他存储装置中。

[0035] 处理装置200包含用于管理处理装置200的操作的操作系统201。在某些实施例中，操作系统201是以软件形式实现的。然而，在其它实施例中，操作系统201可以以纯硬件实现，例如，特别设计的专用电路或部分地以软件和部分作为专用电路来实现。

[0036] 图3是可用于实现任何上述技术的处理装置的框图。请注意，在某些实施例中，至少一些在图3中所示的组件可以在两个或多个物理上独立但相连接的计算平台或箱子之间进行分配。处理装置可以代表传统的服务器级计算机、个人计算机、移动通信装置(例如智能手机)、平板计算机、游戏控制台或任何其它已知的或传统的加工/通信装置。

[0037] 在图3中所示的处理装置301包括一个或多个处理器310，例如中央处理单元(CPU)、存储器320、至少一个诸如以太网适配器和/或无线通信子系统(例如，蜂窝，Wi-Fi，蓝牙，等等)的通信装置340，以及一个或多个I/O装置370、380，其通过互连390全部彼此耦合。

[0038] 处理器310控制处理装置301的操作，并且可以是或者包括一个或多个可编程通用或专用微处理器、微控制器、专用集成电路(ASICs)、可编程逻辑器件(PLD)，或这些装置的组合。互连390可以包括一个或多个总线，直接连接和/或其它类型的物理连接，并且可以包括各种桥接器，控制器和/或适配器，例如公知的现有技术。互连390还可以包括“系统总线”，它可以通过一个或多个适配器连接到一个或多个扩展总线，例如外围组件互连(PCI)总线、超传输或工业标准体系结构(ISA)总线形式、小型计算机系统接口(SCSI)总线、通用串行总线(USB)或电气和电子工程师协会(IEEE)标准1394总线(有时也被称为“火线”)。

[0039] 存储器320可以是或者包括一个或多个类型的一个或多个存储器装置，如只读存储器(ROM)、随机存取存储器(RAM)、快闪存储器、磁盘驱动器等。网络适配器340是适于使所述处理装置301与远程装置通过通信链路传送数据的装置，并且可以是，例如，传统的电话调制解调器、无线调制解调器、数字用户线(DSL)调制解调器、电缆调制解调器、无线收发机、卫星收发信机、以太网适配器，等等。该I/O装置370，380可以包括，例如，诸如一个或多个装置：指点装置，例如鼠标、跟踪球、操纵杆、触摸板，等等；键盘；与语音识别接口的话筒；音频扬声器；显示装置；等等。然而，请注意，这样的I/O装置(只作为服务器工作且不提供直接的用户界面)在系统中可以是不必要的，在至少一些实施例中的服务器通常是这样。在示出的一套组件的其他变型可以与本发明相一致的方式来实现。

[0040] 软件和/或硬件330编程处理器310以执行上述的操作可以被存储在存储器320中。

在某些实施例中，这样的软件或硬件通过从远程系统经由处理装置301(例如，通过网络适配器340)下载被最初提供给处理装置301。

[0041] 图4示出了文本到视频(TTV)系统的示例体系结构。文本到视频系统400可以在单个处理装置或一组处理装置和/或服务器上被执行。该系统400包括含有人的视觉模型的视觉数据库410。该人被称为“目标人物”，“目标个人”，或简称为“目标”，他的脸被称为“目标面”。视觉模型包含人的先验信息，人的先验信息包括，例如，目标人物的图像和视频。在接收到文本430之后，系统400可以创建基于目标人物的视觉模型将文本430的内容映射到目标任务的面部运动的字典。目标人物的视觉序列被基于字典创建。在一些实施例中，来自参考个人的信息也被用于创建视觉序列，如在下列各段中公开的细节。背景场景被在视觉序列中创建，目标面在背景场景之上重叠。

[0042] 系统400还包括包含目标人物的音频模型的音频数据库420。音频模型可以包括目标人物和参考个人的先验信息。有不同的方法来构建音频模型并生成目标人物的音频序列。该方法的详情在下文中被讨论。视觉和音频序列同步，并合并程目标人物(450)的视频序列。在一些实施例中，视频序列被输出在显示器上或输出至远程装置(460)。

[0043] 目标任务的可视化模型

[0044] 图5示出了构建目标人物的视觉模型，尤其是目标人物的面部视觉模型的例子。视觉模型的数据被存储在处理装置上，该处理装置基于文本执行视频生成。在实施例中，视觉模型通过采取人的一个或多个样品视频被创建，而人说出某些字。人所讲的字的数量是需要足够大以使人的面部，嘴唇和嘴的动作被在视频中捕获。在视觉模型创建的阶段，说出的字不必须与视频生成的在后阶段中被提供的字有关系。通常的假设是没有文本的先验知识，该文本的先验知识将要被提供或输入至视频生成。

[0045] 建立视觉模型所需要的是关于嘴和面部的运动的足够信息。在一些实施例子中，即使在不同语言中的样品视频可被用于构建视觉模型。在一个实施例中，例如，所需要的求出数据包含约5分钟的视频。当建立视觉模型时，捕捉典型的面部运动的来自视频的代表帧被选为特征点来构建模型(510)。特征点被手动或自动标记。这些特征点可以包含代表人的重要或典型的面部的点(例如，当上下唇相遇，或上下眼睑相遇时)，以及在重要面部特征之间的中间点。

[0046] 对于每个选定的框架，在框架上的N个点被选择为代表人的面部特征的网格点。因此，每个帧定义了在2N维的欧几里得空间(每个点由x，y坐标表示)中的坐标。由于这些点代表具有不同的情感的个人的脸的形状，在高维空间中不存在随机分散。在一个实施例中，降维方法(如主成分分析(PCA))可以被应用到这些点。例如，类似PCA的线性降维方法可以被应用。关于面部的均值图像的椭圆体被定义，以及主轴被定义为数据自相关矩阵的特征向量。这些主轴根据自相关矩阵的特征向量的特征值幅度被编号。具有最大特征值的特征向量表示在N个点之间的最大变化方向。具有较小特征值的每个特征向量表示不重要的方向变化。在一个实施例中，K最大特征向量是足够用于逼真地表示面部的所有可能的运动。因此，每个面部运动表示为K数字的集合，这些数字被称为乘数。每个乘数表示在K最重要的特征向量之间，从沿相应的特征向量方向上的均值图像偏离的程度。这些特征向量被称为形状特征向量。形状特征向量构成人的形状模型520。在一些实施例中，数目K可基于被处理的视频类型被适应和调整。

[0047] 为了表示在面部中的像素的颜色，均值图像的网格点被用来产生均值图像的三角。三角过程划分面部的图像为多个三角形区域，每个三角区域由三个网格点进行定义。对于从均值图像偏离的任何其他的面部运动，相应的三角测量是基于偏离的网格点(相对于均值图像的网格点)被创建的。在一个实施例中，对于每一个N标记帧，网格点的三角处理被执行。这些三角可以被用来创建从标记的帧中的每个三角区域到均值图像中的对应的三角形区域的线性映射。N标记图像的像素值可以被移动在平均形状的边界内部定义的图像。

[0048] PCA稍后在这些均值图像的区域内定义的图像上被执行。图像的数字在PCA后被保留来表示人脸的纹理。这些保留的图像被称为纹理特征向量。纹理特征向量构成的人的纹理模型。类似于形状特征向量的乘数，纹理特征向量乘数被用来表示面部的像素的颜色(换句话说，纹理)。形状特征向量和纹理特征向量的乘数的集合被用于逼真地重建目标人物的面部540。在一些实施例中，例如，特征向量(或相应的乘数)的总数是约40至50。在转换(处理)装置上，面部运动的每一帧可以通过形状和纹理的特征向量的线性组合被重新创建，使用乘数作为线性系数。在一个实施例中，调整向量被存储在转换装置上。

[0049] 形状模型分割

[0050] 在一些实施例中，目标人物的面部可以分割成多个区域。例如，图6示出了分割目标人物的脸部成两个区域(上部区域和下部区域)的示例性过程。单独的一组形状特征向量(610,620)是用于分别建模上，下区域。关联至形状特征向量(610,620)的独立的一组乘数(614,624)被用于分别基于下部和上部面部形状模型(612,622)独立地建模下部和上部区域。然后，由乘数614所表示的合成的下部区域616与由乘法器624所表示的上部区域626相结合以生成目标人物的合成的全部面部630。为了生成人发表演讲的视频，下部区域可以比上部区域有更大的兴趣。因此，下部区域可以由比上部区域更多的乘数/特征矢量来表示。

[0051] 参考个体的视觉模型

[0052] 在一些实施例中，可以使用上面公开的目标人物的同样的程序建立参考个体的视觉模型。在一个实施例中，采用比目标人物更大的数据集创建参考个体的这些视觉模型，由于这些参考个体的模型将被用于创建将文本内容映射到面部运动的字典。例如，参考个人的视觉模型是通过记录参考个人所讲的一个或多个讲话被创建。讲话的内容是足够大，以为了几乎重现在典型发言中发生的所有可能的具有不同情感的嘴的动作。

[0053] 将文本映射至动作的字典

[0054] 一个或多个字典基于参考个体的视觉模型被创建以用于将文本内容映射到面部运动。在一个实施例中，可能的文本内容被分解为字、音素和发声。发声可以是不能由现有的字来表示的声音。每个字，音素和发声可以有在字典中的至少一个条目。在一些实施例中，字，音素或发声可以有在字典中的多个条目。例如一个字可以有在字典中的对应于不同的情感多个条目。

[0055] 在一个实施例中，在参考个人的视觉模型的创建期间，参考模型以人似乎在生成的视频中所讲的语言读出一些最常见的字。其他字，或具有不同的情感的字，将基于构成音素利用来自参考个人的视频记录的信息被重建。

[0056] 在字典中的每个条目是在字，音素或发声和形状乘数的时间序列(时间序列也被称为帧序列，或在此公开的序列)之间的映射。例如，假设需要将参考个人T视频帧的时间周期来执行说出字“雨”的面部运动，形状乘数的时间序列可以被表示为f(K,T)；k=1至K，t=1

至T。在每帧t中，参考个人的面部运动是由形状乘数f(K,T)的K数表示。因此，总体K*T乘数的集合表示对应字“雨”参考个人的面部运动的序列。因此，在字典中的条目可以是：“雨”：f(K,T)；k=1至K，t=1至T。

[0057] 在一个实施例中，字典中的条目可以使用自动语音识别(ASR)引擎被自动编译。该ASR引擎可以识别字和音素。在一些实施例中，ASR引擎可以进一步识别不是字或音素的发声。如果字“雨”以不同的情感被说出，字典可以包含具有不同情感的字“雨”的多个条目。例如，一个条目目可以是：“雨”+(惊喜的情感)：$f_1(k,t)$，k=1至K，t=1至T。

[0058] 在一些实施例中，一些字是由音素构成。音素的时间序列乘数不仅依赖音素本身，而且也依赖之前和之后说出的相邻音素(或之前或之后的音素的沉默)。因此，字典可以包括音素的多个条目。当条目被用来产生视觉序列时，该音素的条目的选择也依赖于文本序列内提供的相邻音素(或以前或以后的音素沉默)。

[0059] 图7示出了字典条目710的一个例子。该字典条目710将字712映射到乘数的时间序列714。面部动作的每一帧720是由来自乘数714的时间序列的一组乘数所表示的。ASR引擎730可用于编译条目710。

[0060] 通过矢量量化的字典最佳化

[0061] 在一些实施例中，当构建字典时，某些包含在字典中的字和音素可以由参考个体说出很多次。这也产生了包含大量单个字或音素的条目的字典，每个条目将字或音素映射到不同的时间序列乘数。例如，如在之前提到的，在音素的情况，音素字典条目的选择可以在讲话历史的基础上作出(例如，特定音素的口形是依赖于之前或之后说出的相邻音素)。这里的选择可能仍然是太多了，即字典提供了太多的条目选择。

[0062] 为了增加可预测性和检索效率，字典可以以下面的方式进行优化。讲话的视觉序列可以被认为是在乘数值的空间中的非常大数量的点。举例来说，如果以每秒30帧的一个三十分钟视频被使用时，我们有30*1,800*K=54,000*k乘数值的集合，其中K是用于视觉模型的形状特征向量的数目。这些点的一部分表示非常接近彼此的嘴的位置。

[0063] 矢量量化(VQ)可在由在K维空间中的54,000点的集合执行。在矢量量化中，54000点被近似为M质心(VQ点、VQ中心或VQ指数)—其中每个点被替换为它最接近的质心。该中心数字越大，54,000点的VQ点的表示越好。由于脸部的动作代表高度约束点集。在乘数之间具有相关性。因此，积极的VQ表示是可能的。在一个实施例中，VQ中心的数目可以被确定，以使得最大误差是可以容忍的，其中最大的误差能够目视识别以提供可接受的视频性能。

[0064] 因此，矢量量化之后，字典条目将包含VQ中心的时间序列，而不是相应的特征向量的乘数的时间序列。这允许字典的更紧凑的条目来表示字和音素。在原字典中的字或音素的多个条目将很可能“分解”为较少的条目的包含VQ中心的时间序列(因为由参考讲话者在不同的时间作出的相同发声，将有可能映射到VQ中心的同一时间系列)。另外，该分解将基于更易于管理的相邻音素作出选择音素的时间序列的任务。

[0065] 音频模型

[0066] 音频模型的目的是基于给定的文本创建任意句子或句子集合。基于文本创建的音频模型多种技术是将在下面的段落中公开。

[0067] 文本到语音(TTS)的音频型号

[0068] 图8示出了创建TTS声音模型的示例性过程。在TTS中的音频模型中，为了创建一个

基于文本的文件(840)为目标的人的音频模式，语音样本(810)被收集。语音样本中的音频数据被用于创建语音特征为目标的人的集合。在一个实施例中，语音特征包括激励参数(820)和光谱信息(830)。这些语音特征和相应的提取的文本内容(850)是被用于创建和求出音频模型(860,870)的输入。一旦创建了音频模型，新的文本序列可被提供用于生成音频。这个音频模型是一个概率模型，即，给定的新的文本序列，来自音频模型的一组语音特征被组合成最有可能代表新的文本序列的音频序列。

[0069] 例如，图9示出TTS音频合成的示例性过程。文本920被输入到概率模型910。在930，表示语音特征的参数序列被模型选择以表示文本910。表示语音特征的语音参数由模型变换以生成音频波形，以及因此该音频序列被合成(940,950)。

[0070] TTS系统的输出不仅是音频序列，也可以包括字和音素的时间标记(也称为时间戳)。例如，考虑字“雨”是要被转换为音频序列中的文本的一部分。音频模型不仅生成字“雨”，也产生相对于所产生的音频序列的起始时间的字“雨”音频序列的开始和结束时间戳。此时间戳信息可以用于在以后的段落中公开的音视频同步。

[0071] 文本到音频合成的直接TTS模型可以产生直接涉及被用来生成模型的音频数据的语音。这种技术的优点是，一旦建立了模型，语音音频的产生仅需要语音文本。

[0072] 声音转换音频型号

[0073] 用于创建音频模型的另一种技术是基于在两个讲话者之间的对应关系的创建。一个讲话者是“参考”讲话者，另一个是“目标”讲话者。在该技术中，基于相同的文本的语音数据被收集用于目标说话者和参考讲话者。对应关系被建立在参考讲话者和目标讲话者的声波波形之间。这种对应关系基于由参考声音发出的这些新字的音频可以稍后被用来生成目标讲话者的新字的音频。

[0074] 参考和目标声音之间的这种对应关系是以以下方式被建立的。来自目标和参考的人说同样的话的音频样本被收集。在一个实施例中，音频样本有几分钟的长度。根据波形的分析，参考和目标声音的话语对齐，以使得参考和目标声音的话语之间的对应关系可以进行。参考和目标声音两者的声音特征(如梅尔频率倒谱系数)被提取。参考和目标声音之间的联合直方图的特征值分布被创建。此联合分布是由GMM(高斯混合模型)来建模的。GMM的参数的第一个估计值由联合直方图中的特征集群的矢量量化创建的。然后，GMM由EM(期望最大化)算法求出。

[0075] 利用这种技术，参考的声音的特征可以被映射到目标的对应特征。来自这些相应的特征，声音波形被生成作为目标人物的音频序列。在一些实施例中，在该过程的第一步骤中的特征的对准是嘈杂的。所生成的目标声音(与原来的目标声音相反)可以被代入算法作为输入，以迭代地执行它直到达到收敛。

[0076] 这种声音转换模型有一些优势。第一是语音的情感状态将从参考被转移到目标。第二个是，如果参考视频也用于生成语音，它可以帮助更高质量的目标视频呈现。因此，例如，当需要目标的特定和精确的情感影响时，声音转换模型可以有助于娱乐目的。

[0077] 基于声音转换的PCA

[0078] 基本GMM基础语音转换可以通过使用PCA(主成分分析)提高效益和速度。在这种情况下，GMM语音收敛训练被执行用于单个的参考声音和多个目标声音。不同目标声音的多个被求出的GMM可以通过PCA程序进行，这会自然分解声音中的变化。

[0079] 被提供的生成的目标声音样本足够大，增加新的目标声音不再需要多分钟的音频样本采集和新GMM的训练。除了只获得了新目标的短持续时间的语音样本之外，其GMM参数通过分解为PCA特征向量被确定，这是基于以前求出的GMM。根据不同目标的多个GMM的原始源的足够丰富的训练集，生成的声音质量也将得到加强，因为PCA将消除由于单个GMM过程中的噪声的变化。

[0080] 在本技术的一个总结中，参考数据转换为多个训练求出的GMM。为一个PCT模型被生成用于多个求出的目标GMM。对于新的目标人物，PCA分解被进行以合成新目标人物的音频序列，其中只有有限的训练数据被从新的目标人物所要求。

[0081] 基于PCA声音转换的TTS

[0082] 在上一节中提到的参考声音并不一定是自然的人声。它可以是高品质TTS生成的声音。该TTS生成的声音并不必须是特定的个体的声音。如上一节一样的确切的相同的步骤被作出，不同之处在于参考声音是高品质合成TTS，而不是来自参考个人的声音。

[0083] 利用固定的合成TTS源的优点在于，为了产生新的目标的声音，无需返回到一组新字的源音频的生成的人源。因此，只有文本序列被需要作为视频生成的输入。

[0084] 音视频同步

[0085] 所产生的视频和音频序列的同步可以以取决于直接TTS或语音转换方法是否被用来建立音频合成的不同方式来实现。该方法可能需要建立参考个体和目标人物的视觉模型之间的关系。这个关系是由参考和目标的形状特征向量的对齐生成的。该对齐由变换矩阵来表示，该变换矩阵只需要被计算一次并存储在呈现(处理)装置中。这些变换矩阵的大小是很小的。假定目标人物是由20形状特征向量和23纹理特征向量来表示的以及字典是基于由18形状特征向量和25纹理特征向量表示的参考个体被编译。因此，该变换矩阵分别是20x18和23x25的形状特征向量矩阵和纹理特征向量矩阵。为了转换的目的，仅是这些矩阵需要被存储在再现装置上。用于创建字典的参考个体的存储数据库不需要在再现装置上。

[0086] 与声音转换音频型号的同步

[0087] 当目标音频以语音转换方法被生成时，则与视频同步的过程如下。音频基于参考个体被生成，我们为该参考个体同时具有视频和音频模型。参考个体的形状特征值和特征向量的乘数被计算。这些乘数被变换到用于产生目标人物视觉序列的形状特征向量和纹理特征向量的乘数。

[0088] 目标人物的这个生成的视觉序列包含需要与目标人物的音频序列、任何显示在音频序列中的情感同步的面部动作和嘴唇动作。因此，目标人物的视频序列可以通过将音频模型(通过音频转换)和视频模型从参考个体转换到目标人物来实现。情感效果可以通过识别在参考个体的音频和/或视觉数据中的情感来实现。

[0089] 与直接TTS音频型号的同步

[0090] 如果目标的音频用在之前段落中所公开TTS技术生成，则字典映射字和音素到乘数的时间序列被用来实现同步的视频合成。如上面所提到的，对准变换矩阵可以在参考和目标视频模型之间使用。在一个实施例中，如果目标人物是字典所基于的参考个体，则对准变换矩阵不是必需的，并且字典可用于直接对准目标人物的音频和视觉序列。在其它实施例中，不存在基于目标人物的字典。乘数基于参考个体的字典被计算，然后使用一次计算校准变换矩阵将这些乘数转化为目标人物的乘数。

[0091] 图10示出了视频和音频序列之间的同步的示例过程。文本1010被分解成字、音素或话语(1020)。每一个字、音素或话语有一个持续时间1030。每一个字、音素，或话语匹配在字典(1040)中的条目。该条目包含乘数或VQ中心(也称为矢量量化索引)的时间序列。该程序检查由TTS系统生成的音频中的字或音素中的持续时间是否匹配由字典产生的相应的视觉运动。如果持续时间不匹配，这种情况可以通过由TTS系统提供的开始和结束时间戳进行纠正。这些持续时间的比例可以被用来产生与声音(1050)的时间戳相匹配的乘数的时间序列。通过这种方式TTS产生音频序列和字典生成视觉序列之间的同步被实现。

[0092] 因此，参考个体的视觉序列1060的正确同步帧被生成，通过应用参考至目标对准转换(1070)以生成目标人物的视频帧1080。

[0093] 将生成的视频编织到后台

[0094] 上述部分侧重于生成适当的面部动作和目标人物的嘴部运动。对于目标人物的一个完整的视频，身体的其他部分(特别是头发，颈部和肩部)都必须被生成。在本公开中的“背景”可以包括两个区域：(1)不是由视觉模型生成的目标人物的身体部分，及(2)与目标的主体分离的场景。

[0095] 嵌入合成视频为背景的任务是一个分层的过程。场景是由身体部分“覆盖”的以及合成的视频部分填充每一帧的剩余部分。在一个实施例中，不存在视频合成的场景选择的限制。例如，用户可以从菜单中选择所需的场景以达到一定的效果。

[0096] 存在有不是由视觉模型生成的目标的身体部分的更多限制，因为身体部分需要自然地适合目标人物的面部部分。有多种技术来解决在以下段落中公开嵌入的这一部分，并且这些技术可以彼此结合。

[0097] 通过边界拟合误差的最小化的嵌入

[0098] 图11示出了通过最小化边界拟合误差将合成的视频嵌入到背景的示例性程序。对于所生成的目标视频(其现在的视频的视觉部分仅包括目标人物的合成面部部分)的每个帧1110和背景视频被编织在一起。面部部分的边界点的坐标被计算并存储。检索在背景视频中进行，用于背景的最佳区段，其导致了段的边界和合成视频的边界点之间的最小差值。一旦最佳背景段被识别，在背景视频中的最佳帧(1120)被确定，具有合成目标视频的第一帧(1130)的边界误差以该最佳帧被最小化。接着，在所合成的目标视频边界中的点被移动到背景视频。基于内部(目标)坐标和外部(背景)点，形状乘数为目标人物的现实合成被进行调整和重新合成。重复此过程直到边界点都在背景边界点的某些容许误差内。面部合成部分现在被嵌入到非合成部分(1140)之内。面部运动(尤其是嘴部位置)的影响最小，因为来自背景视频的帧被选择为最小化嵌入误差，如上所述。

[0099] 现在进行视频序列的下一个合成帧。相同的边界误差被计算以用于下一合成帧、当前使用的帧和背景视频的之前的帧。在这三个帧中找到一个最小化的边界误差并重复上述的迭代过程以嵌入第二帧。此过程为合成的视频(1150)的每一帧重复。

[0100] 通过分割的嵌入

[0101] 正如之前段落所提，不同的模型可以用于面部的上部和下部。图12示出了使用两个分段模型将合成的视频嵌入到背景的示例过程。在该模型中，面部的上部和下部被装配到目标人物的现有视频中，称为背景视频(1210)。

[0102] 上部边界点是相对刚性的(即在前额)以及通过以简单的刚性变换1220(包括缩

放、改变方向)移动上段中的所有点使之与顶部区域的边界(1230)对准,上面部可以被嵌入到背景中。

[0103] 下段可能不用与上段相同的方式被嵌入到背景中,因为下段的边界点是非刚性(事实上钳口点与讲话相关联并且在边界上)。然而,一些信息仍然可以从合成边界获取。下面部还可以缩放到适当的大小,以便嵌入下面部到背景(1240)。这提供了一个缩放参数以帮助将下段编织到上段。

[0104] 上段和下段稍后以如下方式连接到彼此。下部和上部部分的连接是以这样一种方式,即所述两个部段具有至少3个公共点,被执行的。这些公共点确定如何平移、旋转和缩放下段,以便被连接到上段(1250)。下段是根据公共点(1260)被对齐,以及上、下段被组合以创建要被嵌入的完整的面部(1270)。

[0105] 兴趣区域

[0106] 背景可以被分成多个兴趣区域(ROIs)。例如,如颈部和肩部区域可以被包括在兴趣区域中。合成的视觉序列的边界被跟踪。具有在合成视觉序列和兴趣区域(包括颈部和肩部)之间的最佳匹配的帧可以被选择作为将合成视觉序列嵌入到背景视频的基础。利用兴趣区域的技术在美国专利申请号为13/334,726的申请中被详细讨论,其以引用方式被合并于此。

[0107] 本文所介绍的技术,可以通过以下方式实现,例如,可编程电路(例如,一个或多个微处理器)、编程软件和/或固件,或专用硬件电路中的实体,或者这些形式的组合。特殊用途的硬线电路可以是,例如,一个或多个应用专用集成电路(ASIC)、可编程逻辑器件(PLD)、现场可编程门阵列(FPGA)等形式。

[0108] 对于在实施此处提出的技术中使用的软件或硬件可以被存储在机器可读存储介质上,并且可以由一个或多个通用或专用的可编程微处理器执行。“机器可读存储介质”,如在本文中使用的术语,包括任何可以以机器可访问的形式存储信息的机制(该机器可以是,例如,计算机、网络装置、蜂窝电话、个人数字助理(PDA)、制造工具、任何具有一个或多个处理器的装置,等等)。例如,机器可访问存储介质包括可记录/不可记录介质(例如,只读存储器(ROM)、随机存取存储器(RAM)、磁盘存储介质、光存储介质、闪存装置),等等。

[0109] 术语“逻辑”,如本文使用的,可包括,例如,用特定的软件和/或硬件进行编程的可编程电路、专用硬件电路,或它们的组合。

[0110] 除了上述实施例,本发明的各种其它修改以及改变可以在不背离本发明的前提下被作出。因此,上述公开并不应被认为是限制性的,所附权利要求被解释为包含本发明的真实精神和整个范围。

[0111] 在一个实施例中,一种方法被引入。该方法包括在处理装置输入文本序列;以及由处理装置基于文本序列生成人的视频序列以模拟人的视觉和听觉情感表达,包括使用人的声音的音频模型生成视频序列的音频部分。

[0112] 在相关实施例中,处理装置是移动装置,文本序列从第二移动装置通过短消息服务(SMS)信道被输入,以及生成人的视频序列包括由移动装置基于存储在移动装置和第二移动装置上的共享信息生成人的视频序列。

[0113] 在另一相关实施例中,文本序列包括一组字,一组字包括至少一个字,以及其中视频序列被生成,以使得人似乎在视频序列中说出字。

[0114] 在另一相关实施例中，文本序列包括表示发声的文本，以及其中视频序列被生成产生，以使得人似乎在视频序列中说出发声。

[0115] 在另一相关实施例中，文本序列包含字和字的指标，当人似乎在视频序列中说出字时，该指标指示在视频序列中的一个时间上的人的情感表达，指标是在预定的指标集合之内，并且预定的指标集合中的每个指标是与不同的情感表达相关联。

[0116] 在另一相关实施例中，生成视频序列包括由处理装置生成人的视频序列以基于文本序列和人的先验知识模拟人的视觉和听觉情感表达。

[0117] 在另一个相关的实施方案中，先验知识包括该人的照片或视频。

[0118] 在另一相关实施例中，生成视频序列包括映射在文本序列中的字至人的面部特征；以及在背景场景中呈现人的面部特征。

[0119] 在另一相关实施例中，字基于字的一个或更多个指标被映射到面部特征，其中所当人似乎在视频序列中说出字时，指标指示在视频序列中的一个时间上的人的情感表达。

[0120] 在另一个相关实施例中，面部特征包括适用于多个人的一般面部特征。

[0121] 在另一相关实施例中，面部特征包括专门适用于人的特定面部特征。

[0122] 在另一相关实施例中，生成视频序列还包括：生成与人的面部特征相适应的人的身体姿势。

[0123] 在另一相关实施例中，生成视频序列包括通过使用基于人的声音的音频模型，基于文本序列中的字生成表示人的说话的音频序列。

[0124] 在另一相关实施例中，文本序列的接收包括实时接收文本序列；其中视频序列的生成包括基于文本序列实时生成人的视频序列以模拟人的视觉和听觉情感表达，包括使用人的声音的音频模型来生成视频序列的音频部分。

[0125] 在另一相关实施例中，另一种方法被引入。该方法包括在处理装置输入文本序列；由处理装置基于文本序列生成人的视觉序列以模拟人的视觉情感表达，其中视觉序列的每个帧的面部是由人的先验图像的组合来表示的；由处理装置基于文本序列生成人的音频序列以模拟人的发声情感表达，利用人的声音的声音模型，以及由处理装置基于文本序列通过合并视觉序列和音频序列制造人的视频序列，其中基于文本序列，视觉序列和音频序列是同步的。

[0126] 在另一相关实施例中，视觉序列的每个帧的面部是由人的先验图象的线性组合所表示，以及人的先验图像的每一个先验图像对应于人的均值图像偏差。

[0127] 在另一相关实施例中，生成基于文本序列的人的视觉序列包括将视觉序列的每个帧划分成两个或多个区域，其中至少一个的区域是由人的先验图像的组合来表示。

[0128] 在另一相关实施例中，人的声音的音频模型包括多个从人的语音样本中创建的声音特征，多个声音特征中的每个声音特征对应于一文本。

[0129] 在另一相关实施例中，多个声音特征中的每个声音特征对应于字，音素，或发声。

[0130] 在另一相关实施例中，人的声音的音频模型包括多个声音特征，声音特征从人的语音样本、第二人的根据文本序列的语音，以及在人的声音的波形与第二人的声音波形之间的对应关系被创建；以及其中基于人的声音波形和第二人的声音波形之间的对应关系人的声音特征被映射到第二人的语音。

[0131] 在另一相关实施例中，人的声音的音频模型包括多个声音特征，声音特征从人的

语音样本，根据文本序列由文本-语音模型所产生的语音，以及人的声音波形和文本-语音模型的声音波形之间的对应关系创建，以及其中基于人的声音波形和文本-语音模型的声音波形之间的对应关系人的声音特征被映射到语音。

[0132] 在另一相关实施例中，另一种方法被引入。该方法包括创建文本序列，其中文本序列代表一个或多个字，一个人将要在视频序列中说出一个或多个字，视频序列使用基于人的声音的音频模型被生成，视频序列视觉地和听觉地表示人的情感表达范围；识别与文本序列中的字相关联的指标，其中指标是在预定指标集合中的一个，其中的每个指示人的不同的情感表达；将指标并入到文本序列；以及发送文本序列至被配置为生成视频序列的装置。

[0133] 在另一相关实施例中，确定识别指标包括从条目菜单中选择一个条目与文本序列中的字相关联，其中在菜单中的每个条目是表面人的情感表达的指标。

[0134] 在另一个相关实施方案中，识别指标包括：将标记语言字符串与文本序列内的字相关联，其中标记语言的字符串来自预定的标记语言字符串的集合，以及在预定的标记语言字符串的集合中的每个标记语言字符串是表明该人的情感表达。

[0135] 在另一相关实施例中，识别指标包括使用自动语音识别(ASR)引擎基于讲话者读出所述文本序列中的所述字的音频序列识别文本序列中的字相关联的指标。

[0136] 在另一个相关实施方案中，说话者是一个不同的人从说的人。

[0137] 在另一相关实施例中，另一种方法被引入。该方法包括在处理装置中存储多个非人物体的先验信息；以及基于存在于处理装置中的多个非人物体的先验信息生成多个非人物体的视频序列，其中多个非人物体的每一个是独立可控的。

[0138] 在另一相关实施例中，多个非人物体被限制于在视频序列中的其他元素。

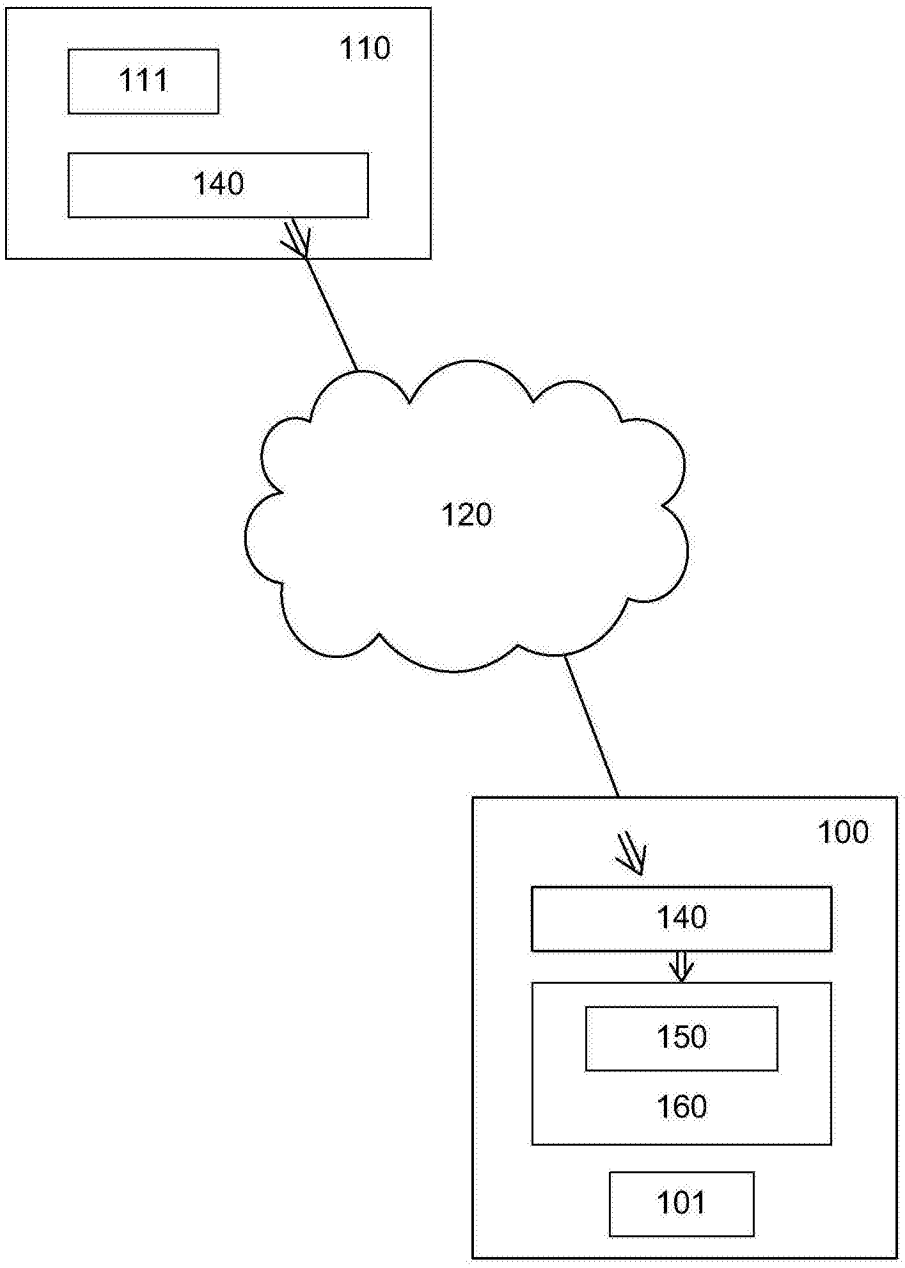


图1A

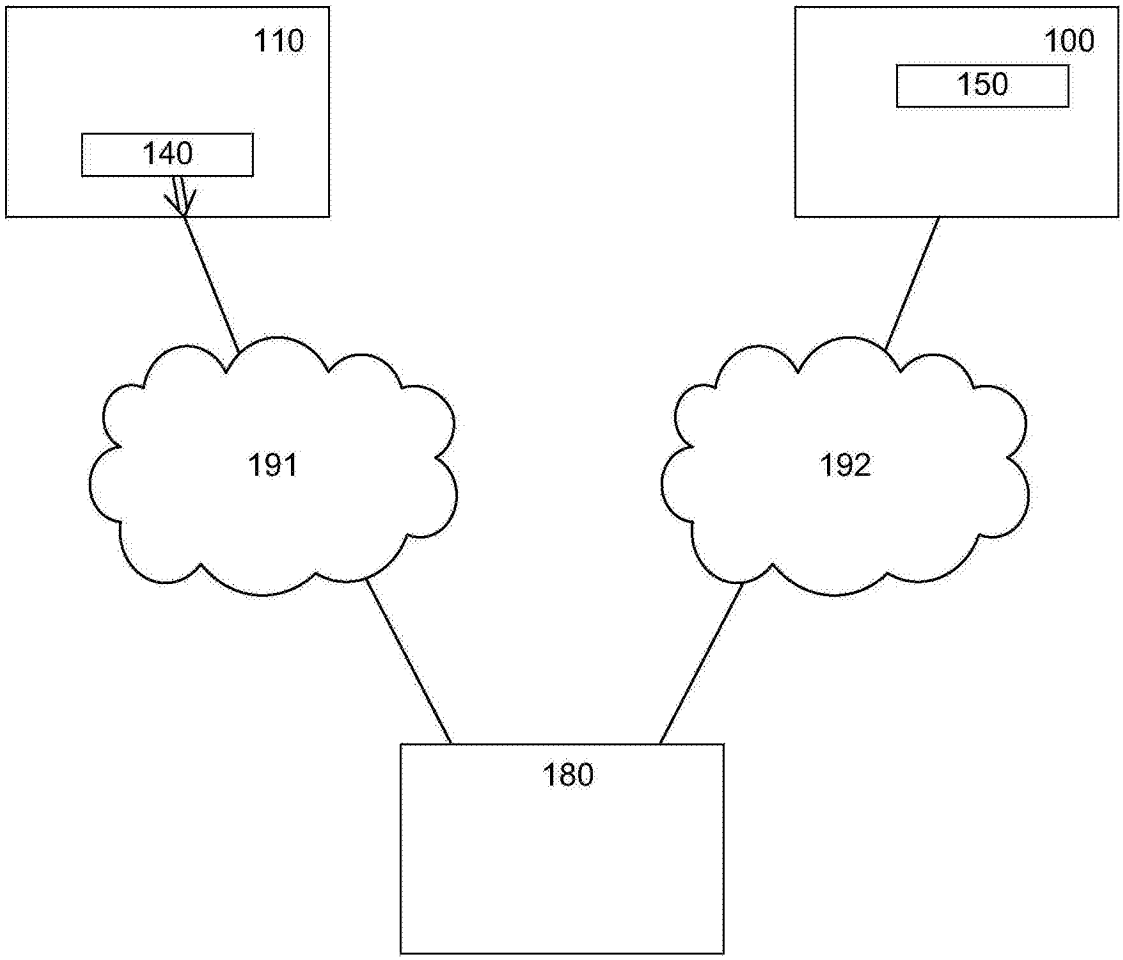


图1B

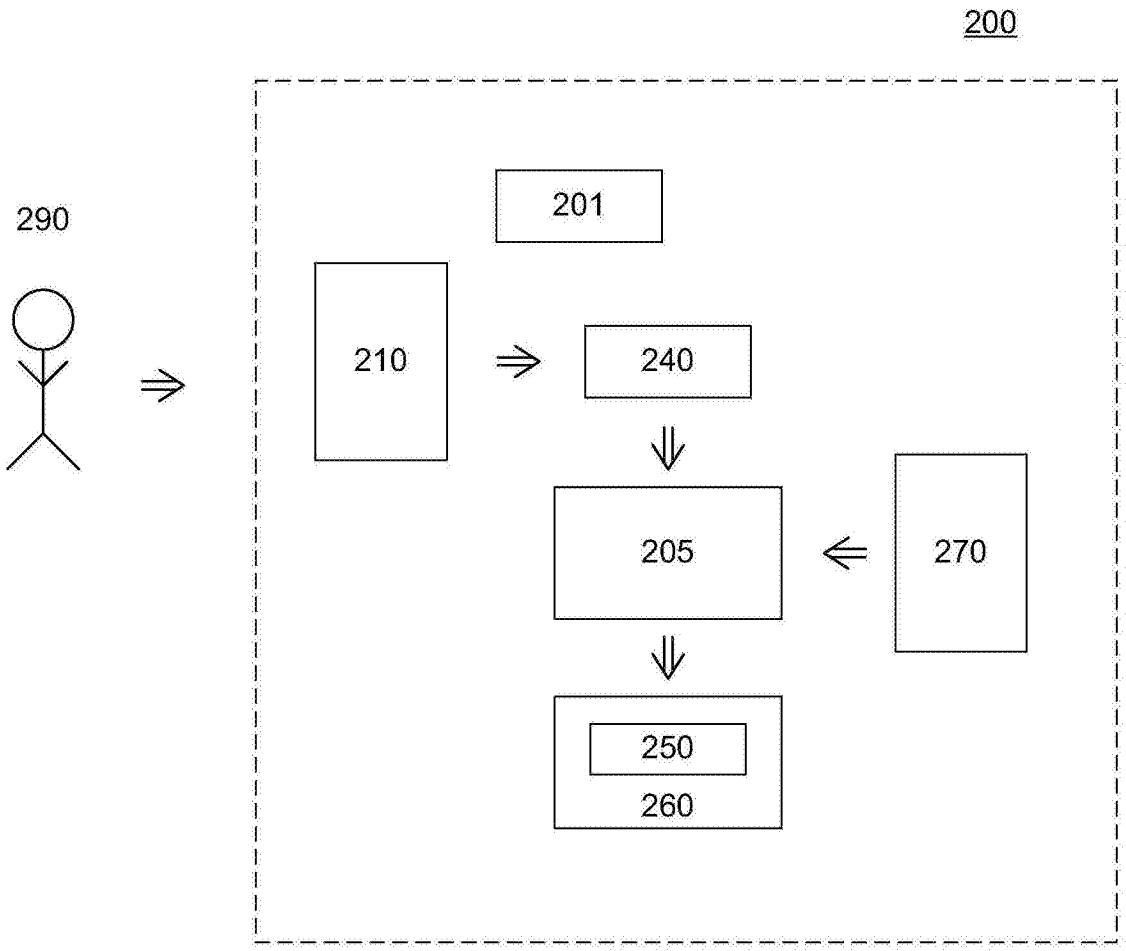


图2

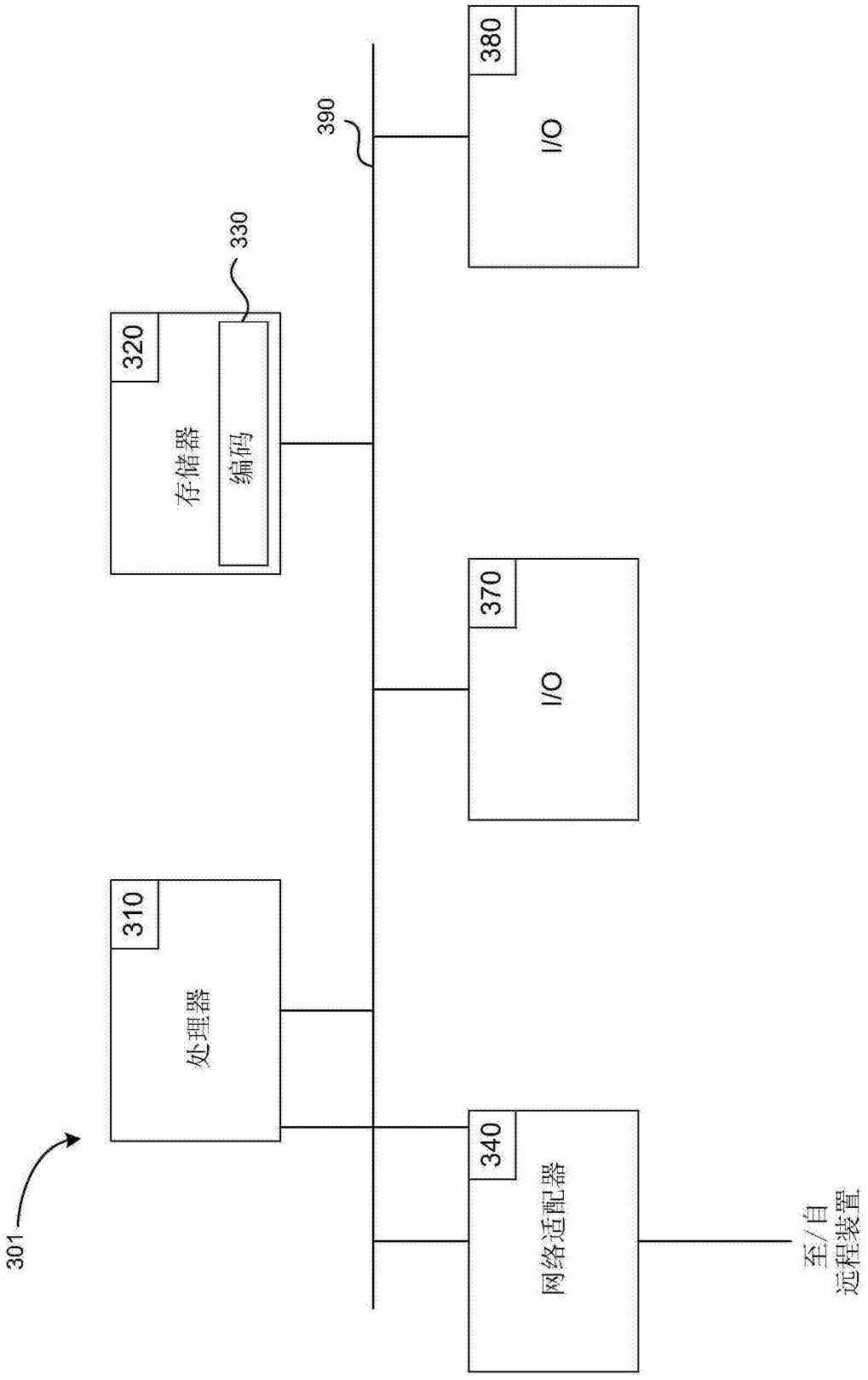


图3

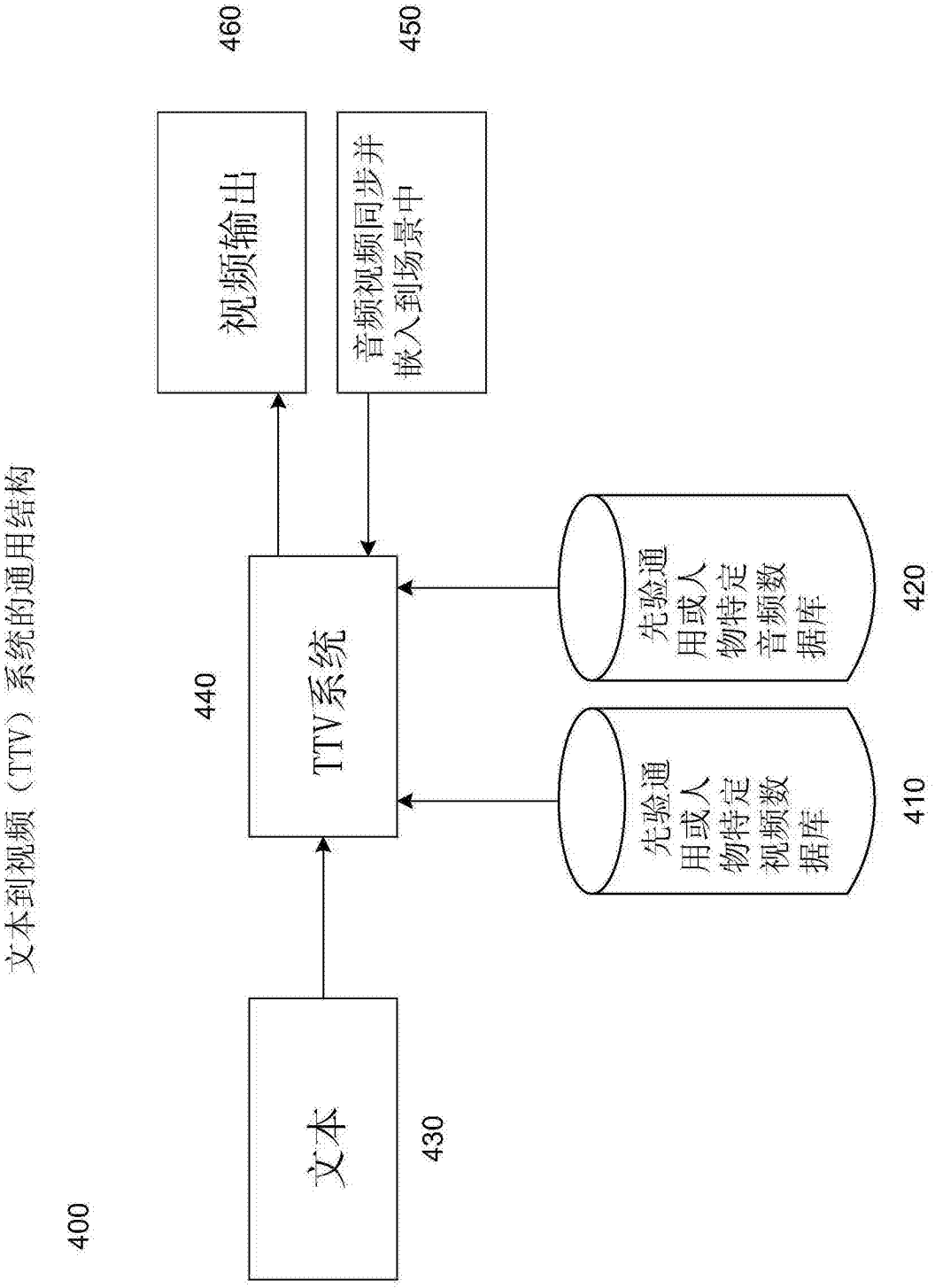

文本到视频（TTV）系统的通用结构
400
文本
430
TTV系统
440
先验通用或人物特定视频数据库
410
先验通用或人物特定音频数据库
420
音频视频同步并嵌入到场景中
450
视频输出
460

图4

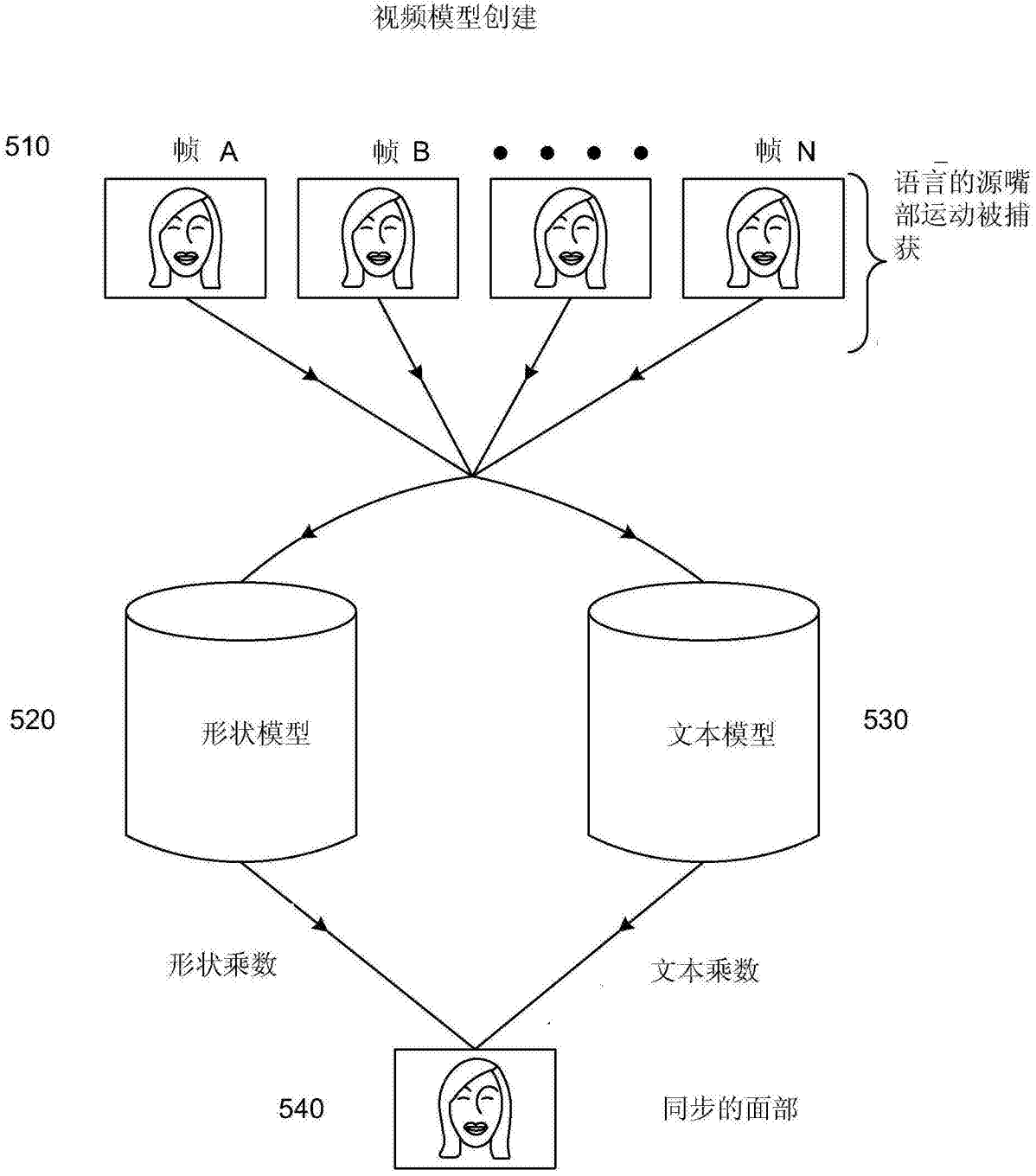

视频模型创建
510
帧 A
帧 B
帧 N
语言的源嘴部运动被捕获
520
形状模型
530
文本模型
形状乘数
文本乘数
540
同步的面部

图5

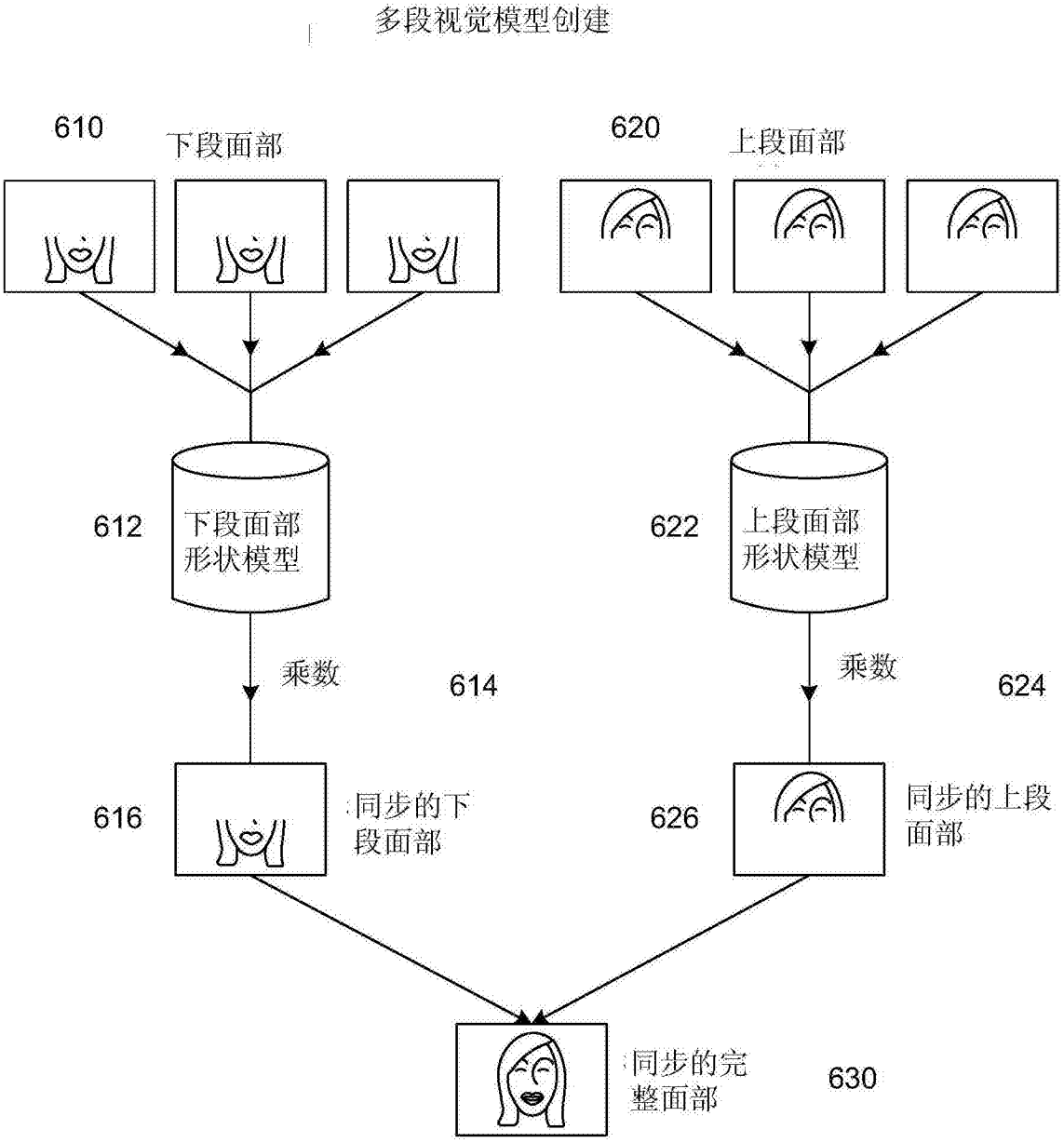

多段视觉模型创建
610
下段面部
620
上段面部
612
下段面部
形状模型
622
上段面部
形状模型
乘数
614
乘数
624
616
同步的下
段面部
626
同步的上段
面部
同步的完
整面部
630

图6

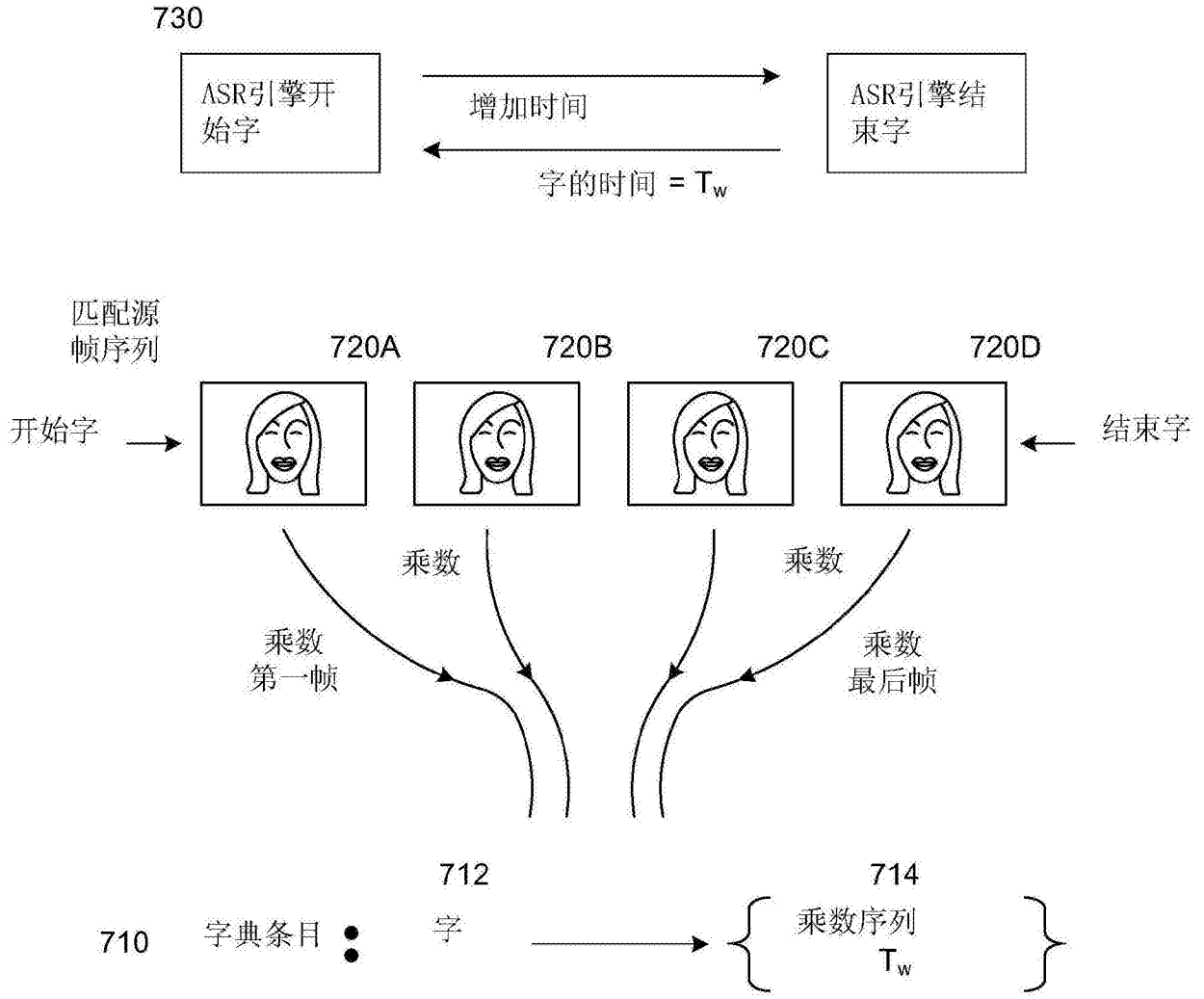

文本至嘴部运动字典创建
730
ASR引擎开始字
增加时间
字的时间 = T_w
ASR引擎结束字
匹配源帧序列
720A
720B
720C
720D
开始字
结束字
乘数
乘数
乘数第一帧
乘数最后帧
712
714
710
字典条目
字
乘数序列 T_w

图7

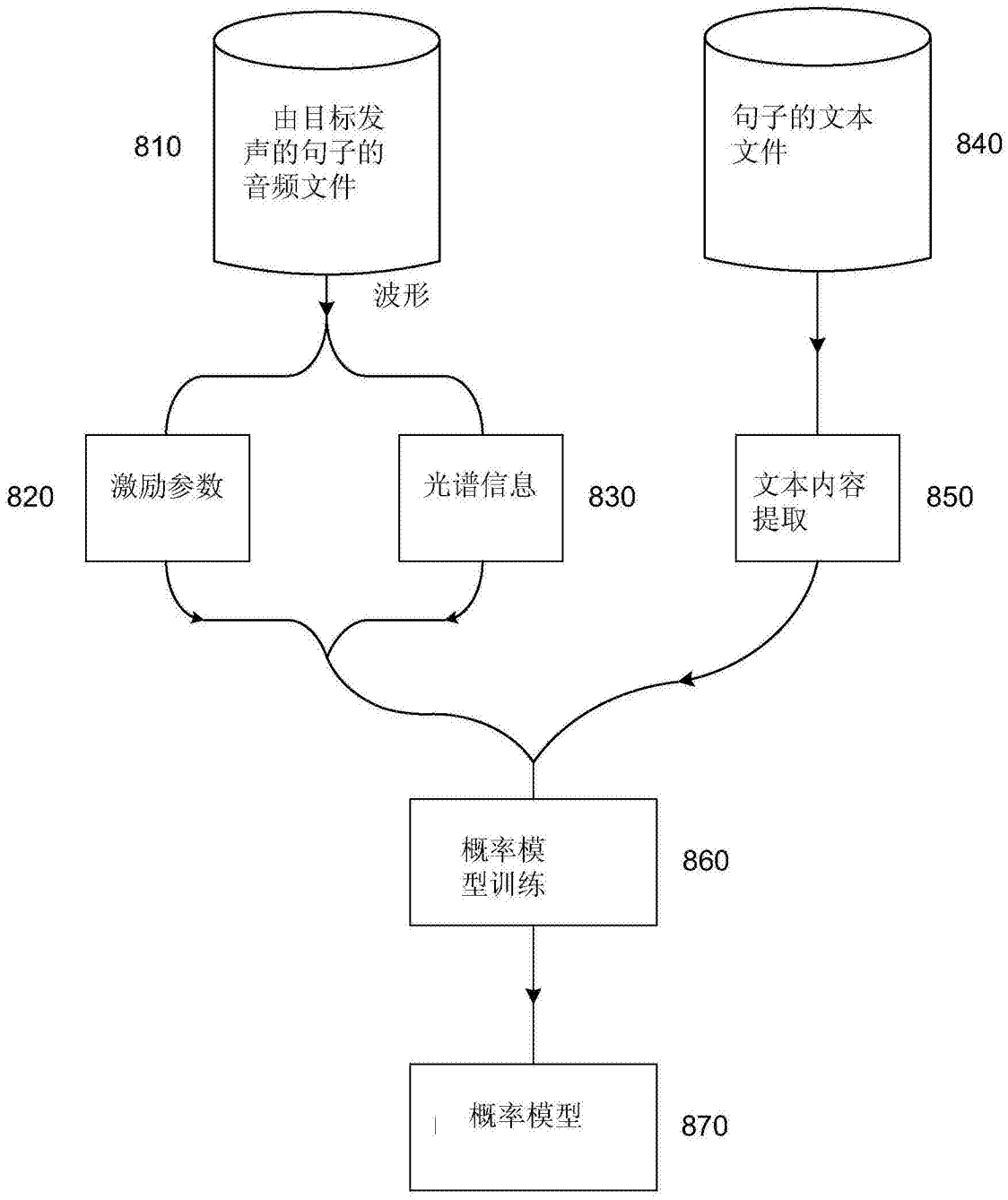

TTS音频模型创建
由目标发声的句子的音频文件
810
句子的文本文件
840
波形
激励参数
820
光谱信息
830
文本内容提取
850
概率模型训练
860
概率模型
870

图8

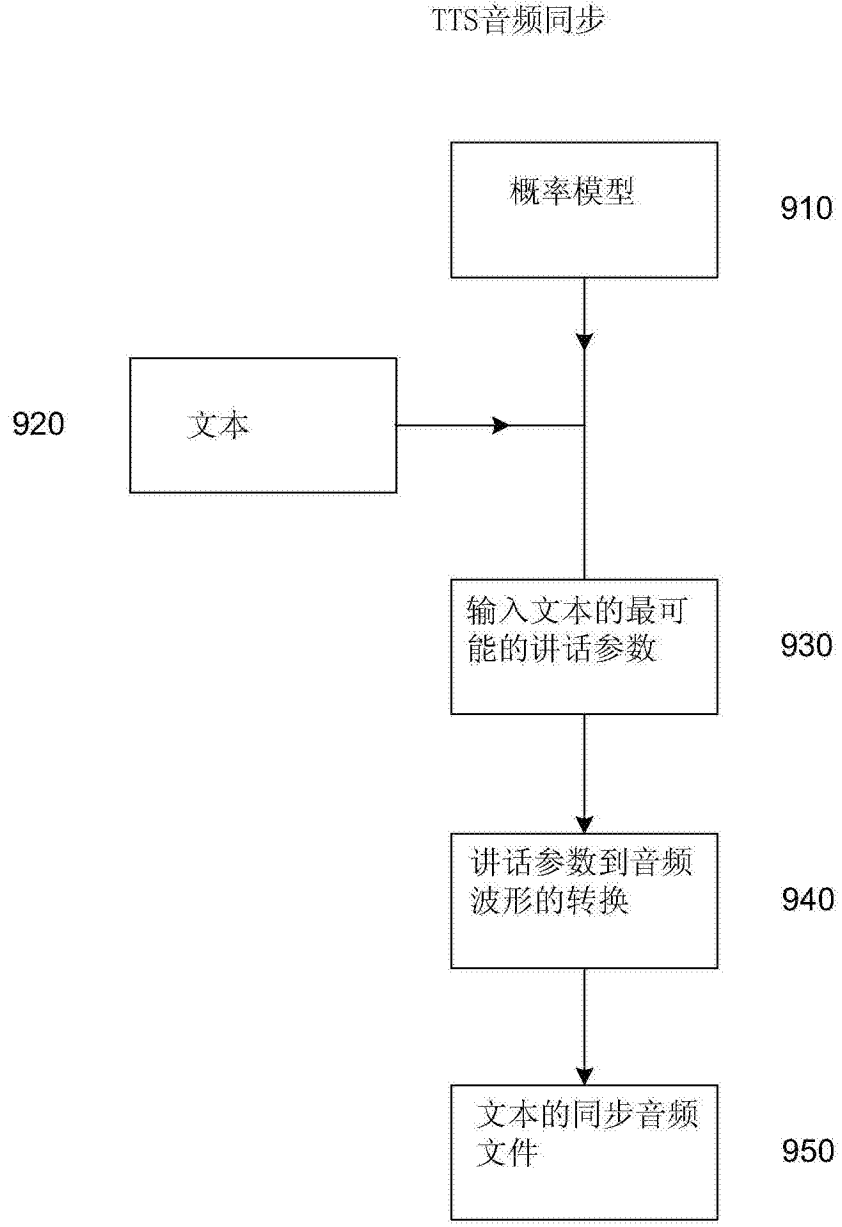

TTS音频同步
概率模型
910
920
文本
输入文本的最可能的讲话参数
930
讲话参数到音频波形的转换
940
文本的同步音频文件
950

图9

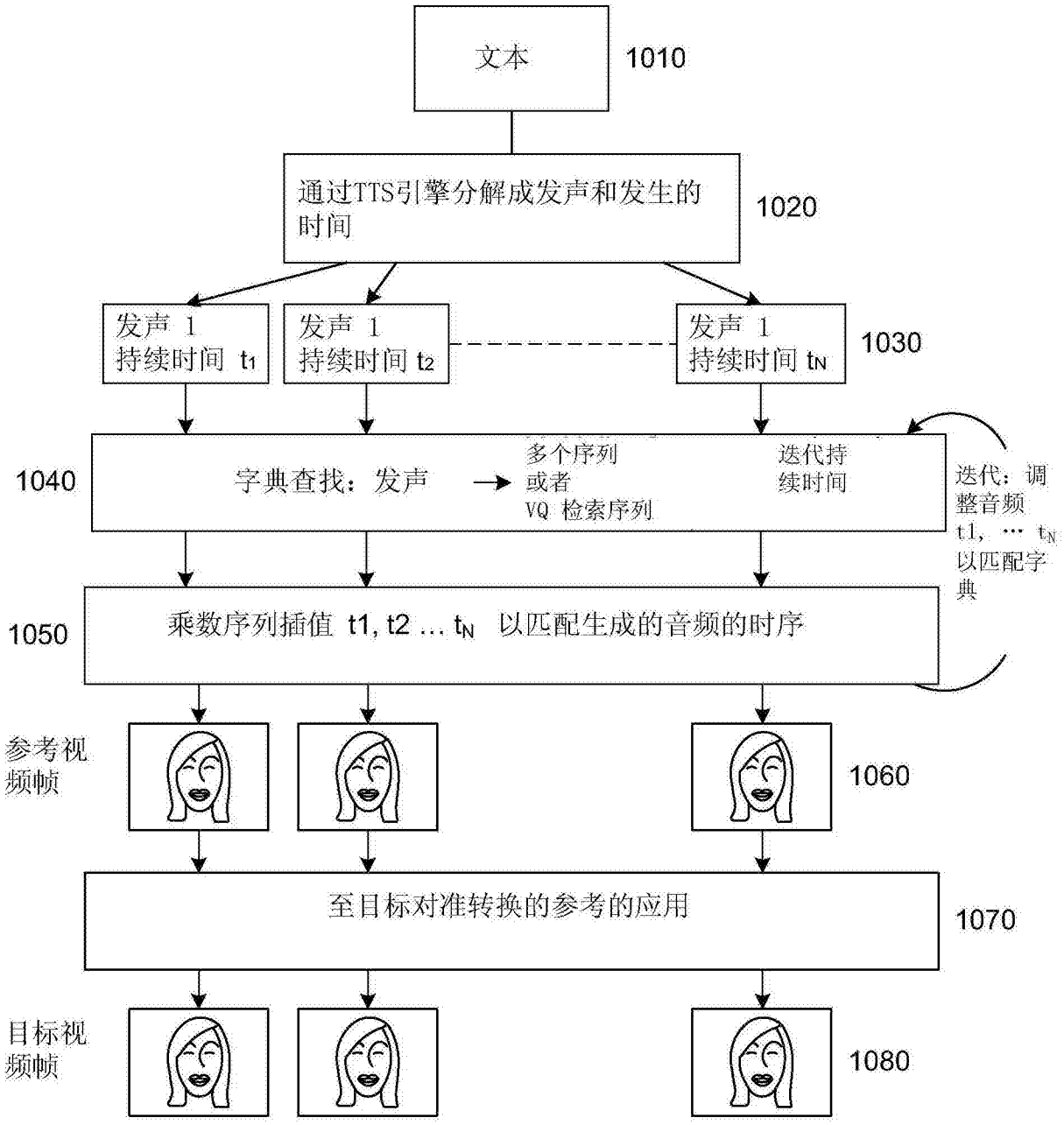

文本
1010
通过TTS引擎分解成发声和发生的时间
1020
发声 1 持续时间 t1
发声 1 持续时间 t2
发声 1 持续时间 tN
1030
1040
字典查找：发声
多个序列 或者 VQ 检索序列
迭代持续时间
迭代：调整音频 t1, … tN 以匹配字典
1050
乘数序列插值 t1, t2 … tN 以匹配生成的音频的时序
参考视频帧
1060
至目标对准转换的参考的应用
1070
目标视频帧
1080

图10

通过边界误差最小化匹配至场景

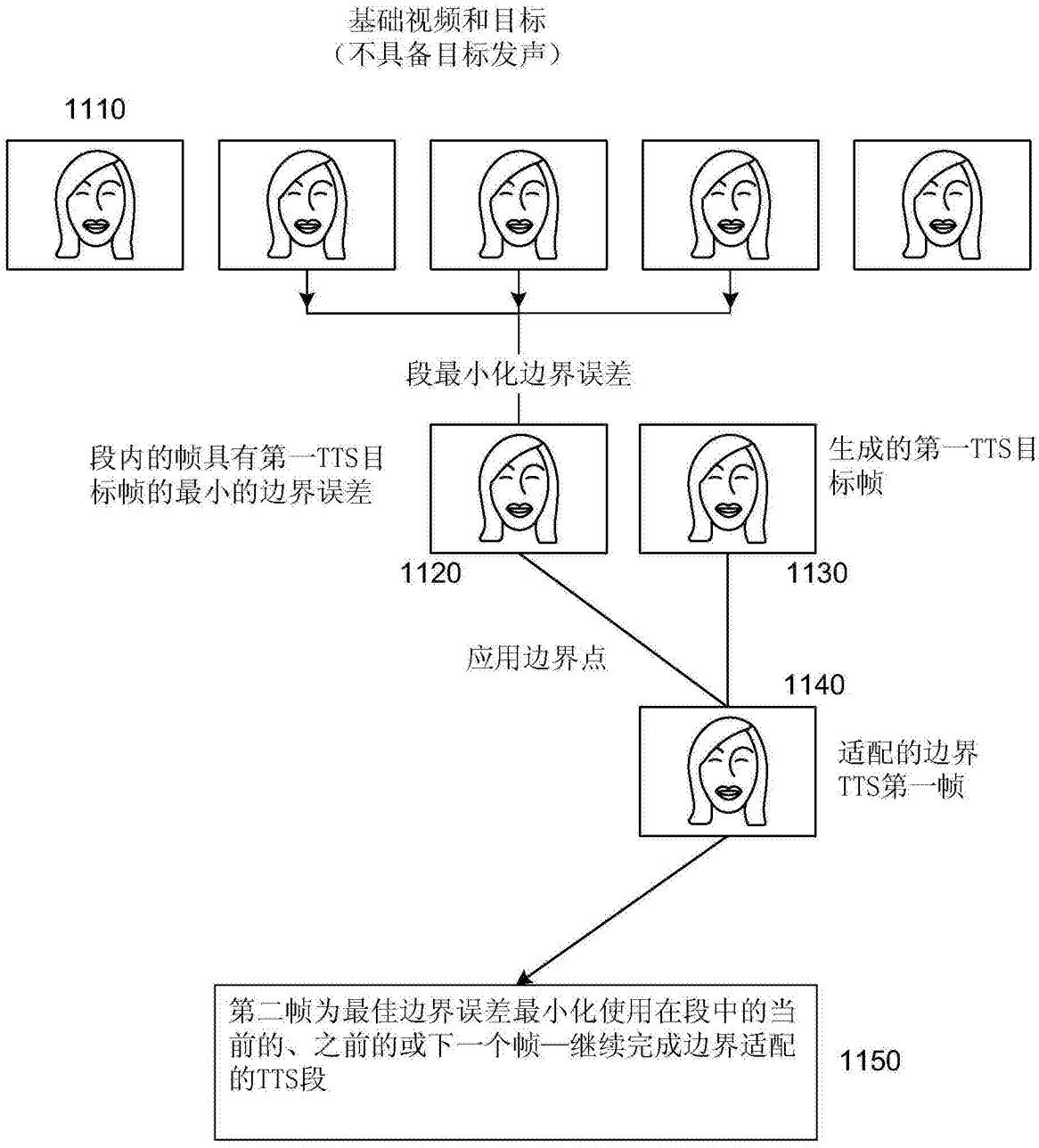


图11

使用分隔的上段和下段面部模型匹配至场景

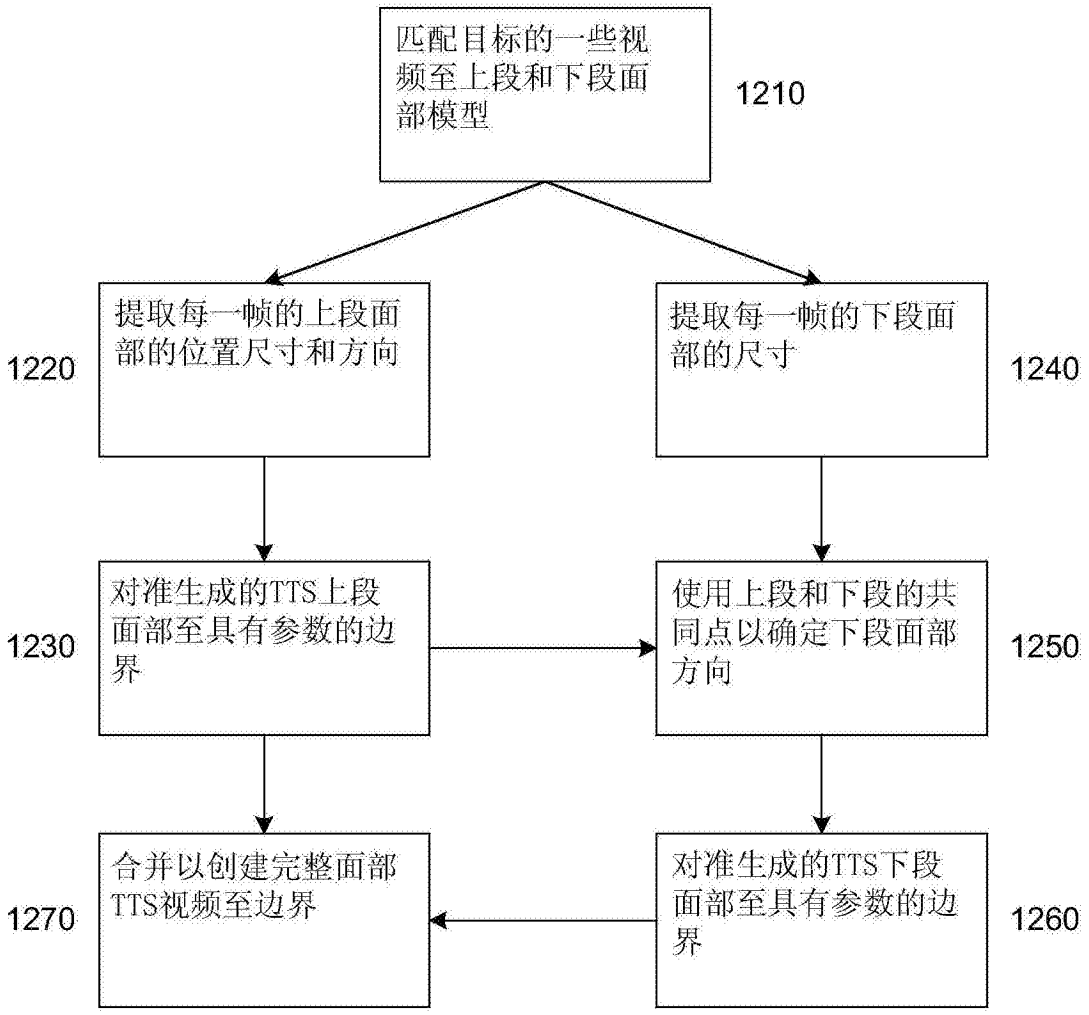


图12