

基于可视语音合成的 3D 通信技术研究 *

杨志晓, 隋 菲, 张德贤

(河南工业大学 信息科学与工程学院, 郑州 450001)

摘 要: 提出了基于可视语音合成的 3D 通信技术概念。通信双方利用文本交换信息, 用户终端采用可视语音合成技术对接收到的文字进行人物化朗读; 同时通过 3D 用户替身的肢体动作和面部表情表达文字蕴涵的人类情感和意图信息, 给出了基于可视语音合成的通信平台结构、情感和意图的表达模型、用户替身的自主交互模型。采用程序驱动方法, 通过控制虚拟人各连杆间沿关节自由度方向夹角及角度变化率等特征参数实现肢体运动合成; 利用 VB 6.0 和 OpenGL 实现了基于可视语音合成的 3D 通信平台原型。

关键词: 可视语音合成; 3D 通信; 用户替身; 非语言信息; 界面

中图分类号: TP391.41; TP393 **文献标志码:** A **文章编号:** 1001-3695(2009)11-4209-03

doi:10.3969/j.issn.1001-3695.2009.11.058

Research on 3D communication technologies based on text-to-visual speeches

YANG Zhi-xiao, SUI Fei, ZHANG De-xian

(College of Information Science & Engineering, Henan Industry University, Zhengzhou 450001, China)

Abstract: This paper proposed a concept of 3D communication technology based on text-to-visual speeches (TTVS). Partners of a communication pair exchanged messages with text. At each end interface, adopted TTVS technology to read received text message. Simultaneously, a 3D avatar expressed embedded human emotions and intentions with body languages and facial expressions. Proposed the structure of such a communication platform. Analyzed expression models of human emotions and intentions, and proposed autonomous interaction models of avatar. Driven by program, obtained avatar's movement by controlling angles and their varying rates between neighbored links. Based on the proposed methods, developed a prototype of 3D communication platform with function of TTVS under VB 6.0 and OpenGL.

Key words: text-to-visual speeches; 3D communication; avatar; nonverbal information; interface

可视语音合成技术通过虚拟人(脸)实现文字的人物化朗读, 在将文字合成为语音进行听觉增强的同时, 通过虚拟人表情脸像与肢体动作的合成来表达一定的情感、意图等非语言信息, 进行视觉增强。由于可视语音合成技术充分利用了文本文件存储容量小、占用带宽资源少、传输速度快等优点, 同时通过语音和视觉增强手段形象表达文字承载的信息, 近年来可视语音合成的研究正逐步形成一个热点。Kurose 等人^[1]研究了基于三维虚拟人脸的日文可视语音合成技术; 王志明等人^[2]采用数据驱动方法实现中文的可视语音合成; 严宝才等人^[3]研究了基于两层隐马尔可夫模型的可视语音合成方法; 杨志晓等人^[4,5]提出了基于肢体动作和面部表情的多模态中文可视语音合成方法; Zhe xu 等人^[6]研究了基于特征词汇的英文文本中人类情感搜索引擎等。文语合成、可视语音合成等技术的发展使其在以文本为存储和传输媒介的各类系统中的应用成为可能。

互联网、无线通信等技术的发展促成了大量即时通信应用, 如 MSN、ICQ、QQ、雅虎通、网易 POPO、手机短信息等, 这些通信平台一般能够提供文本、语音、视频等通信手段。语音和视频通信相对于文本通信而言, 占用较多的带宽资源, 容易出现传输延迟、数据丢包等问题, 导致较低的通信质量。相对而

言, 文本文件在存储和传输过程中具有明显的优势, 因此文本依然是各种即时通信软件的主流信息交换媒介。然而, 当前主要的即时通信工具一般将文字直接呈现给用户, 用户必须用眼睛阅读文字获取信息。长时间的使用眼睛容易造成视觉疲劳甚至影响视力, 而阅读文字将限制用户的其他活动。因此有必要在充分保留文本文件在存储和传输中的优势的同时, 研究新的即时通信技术以提高信息传递效率, 降低用户使用软件的劳动强度。

针对上述需求, 本文研究一种基于可视语音合成的新型文本即时通信技术, 尝试利用 3D 虚拟人对接收到的文字进行人物化朗读, 实现文本通信的语音增强和视觉增强, 以构建高效、逼真、形象的 3D 通信平台。

1 3D 通信技术构架

通信平台结构可以用图 1 来表示。用户 A 和 B 通过服务器建立点对点连接。在各个用户终端, 为通信对另一方的参与者设置用户替身 (avatar), 形成用户—(对方) 用户替身交互场景。图 1 中, 终端 A、B 的交互过程可以表示为: $A \leftrightarrow B' \leftarrow B$ 和 $B \leftrightarrow A' \leftarrow A$ 。通信双方通过用户终端经互联网发送和接收文本

收稿日期: 2009-01-17; **修回日期:** 2009-03-26 **基金项目:** 河南省重点科技攻关计划资助项目 (082102210096); 河南省教育厅自然科学研究计划资助项目 (2008A520006)

作者简介: 杨志晓 (1974-), 男, 河南郑州人, 副教授, 博士, 主要研究方向为人机交互与虚拟现实、可视语音合成、智能技术、理论与系统; 隋菲 (1986-), 女, 硕士研究生, 主要研究方向为人机交互与虚拟现实、可视语音合成 (suifei2000@163.com); 张/ 贤 (1961-), 男, 河南郑州人, 教授, 博士, 研究方向为数据挖掘、智能计算。

信息。接收到的文字在终端用户界面被合成为语音,同时用户替身伴以适当的面部表情和肢体动作来表达文本信息发送者的情感、意图等状态。

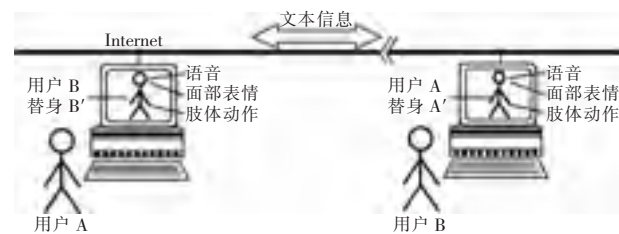


图 1 基于可视语音合成的 3D 通信平台结构

终端用户界面的结构及可视语音合成过程(图 2)可以分为三个层次:a)最高层。接收到的文字被送入非语言识别模块,在本地对文字中蕴涵的情感、意图等非语言信息进行识别;或者由远程通信者在发送端人工指定非语言信息类别。b)中间层。检索非语言信息表达模型库,对识别或指定的非语言信息类别匹配以适当的面部表情和肢体动作类型;同时由用户替身自主交互模型库来决定在用户未作出任何回应时其替身应表现的面部表情和肢体动作类别。c)最底层。由文语合成工具实现语音合成,同时通过面部表情和肢体运动控制模块实现 3D 替身的面部表情和肢体动作合成,完成 3D 渲染,最终通过用户替身将文本信息及其蕴涵的非语言信息以语音、面部表情和肢体动作形象表达。

图 2 中由信息发送者的操作产生的其替身面部表情和肢体动作的过程为受(信息发送者)控交互。事实上,在一定条件下,用户替身还需在其所代表的用户没有作出反应时对通信对的另一方作出反应,表现出一定的智能交互行为。如用户 B 长时间没有键入文字对 A 作出回应,B 的替身应该以适当的方式表达 B 的状态以便 A 了解,从而使得交互过程更加接近真实。这种在用户没有作出反应时,由其替身自主地对另一用户作出反应的过程可称为用户替身的自主交互。因此图 2 中还需有一个自主交互模型,与受控交互过程一起构成与用户的人机交互。

2 非语言信息识别

2.1 非语言信息的分类模型

本文考虑人类情感和意图两类非语言信息。情感和意图都应被识别并以适当的语调、面部表情和身体语言形象表示。

人类感情分为八大类,即

emotion = [高兴 愤怒 伤心 害怕 厌恶 吃惊 讽刺 痛苦]

每种感情有其自己的强度等级,在其表达模型中要设定调节因子,以对应感情的强烈程度。

对人的意图,本文将其分为以下类别,即

intention = [自我介绍 问候 祝贺 求助 帮助 正面评价 反面评价 怀疑 赞扬 抱怨 道歉 感谢 要求 祝愿 想念 再见]

同样,每种意图也存在强度等级,也需在其表达模型中设定调节因子,以对应意图的强烈程度。

2.2 非语言信息的识别

非语言信息识别解决给定文本中人类情感、意图有无的判断、非语言信息分类、强度定级等问题,其高级阶段为智能识别技术。当前研究一般通过搜索特征词汇的方法识别非语言信息,并进行强度定级^[6]。但是识别效果受词的歧义性限制,并

且难以应对大量新兴的所谓互联网语言(如 U2 表示 You too, CU 表示 See you 等)。

本文采用以人工指定为主、人工指定与自动识别相结合的方式给出非语言信息类别。在大量情况下,由文本信息发送者在发送信息的同时指定其情感或意图类别;接收端根据接收到的非语言信息类别对其表达模型进行匹配。对简短的文本信息,建立其非语言信息类型知识库。知识库采用大量“if...then...”语句描述文本信息及其对应的非语言信息类别。如

if message = “高兴” then emotion = “高兴”
if message = “我很高兴” then emotion = “高兴”
if message = “你好” then intention = “问候”
if message = “你也好” then intention = “问候”
if message = “拜拜” then intention = “再见”
if message = “CU” then intention = “再见”
.....

如果信息发送端没有指定其非语言信息类别,接收端将接收到的文本信息送入非语言信息类型知识库中进行自动识别。知识库在一定程度上能够满足非语言信息的智能识别需求。

3 用户替身的交互模型

3.1 非语言信息的表达模型

非语言信息表达模型描述以什么样的语调、面部表情和肢体动作来表达人类特定情感或意图。根据专家经验知识归纳各种情感、意图的表达模型,以情感“高兴”为例,表 1 列出了其语义描述的表达模型。

表 1 感情“高兴”的表达模型

项目	等级				
	1	2	3	4	5
语调	中	中	高	高	高
面部表情	微笑	微笑	中笑	中笑	大笑
身体语言	无	摇头晃脑	挥手	站立、挥手	站立、举手、转动身体

设置特征参数,将表 1 所示语义描述的非语言信息表达模型量化,便于后期的肢体运动合成。

非语言信息表达模型中需设调节因子,以对应情感、意图的强烈程度。

3.2 用户替身的自主交互模型

理想情况下,用户替身的行为应该是人类真实交互行为的移植和再现。用户替身除对其所代表的用户在回应(如发送信息)通信对的另一方时要表现其行为外,还应在用户未作出回应期间自主地表现出一定的智能交互行为,以增加虚拟场景的真实感,便于通信对的另一方了解该用户的真实状态。例如,图 1 中用户 B 如果长时间没有对 A 作出回应,则可认为 B 对 A 的谈话内容不感兴趣,此时 B' 应表现为惰性、睡意等状态。因此,用户替身的行为由其所代表的用户和自主交互模型共同决定。对图 1 中的用户 A、B 而言,交互过程可以描述为

$$A \leftrightarrow B \left\{ \begin{array}{l} \leftarrow B \\ \leftarrow I_{\text{auto}} \end{array} \right. \text{ 和 } B \leftrightarrow A \left\{ \begin{array}{l} \leftarrow A \\ \leftarrow I_{\text{auto}} \end{array} \right.$$

其中 I_{auto} 为用户替身的自主交互模型库。

为提高虚拟交互场景的真实感,需要发掘用户替身自主交互行为的种类和判定方法。本文以距用户上次操作在时间序列上的间隔 T 为特征值,根据专家经验知识判断用户当前状态,决定用户替身行为。

4 文语合成

文语合成实现文字到语音的转换。目前世界上已有多种文语技术能够将对应的语言文字合成为语音,许多还支持双语种混读。中国科技大学讯飞公司开发的中文文语合成 TTS SDK 代表了世界上中文文语合成的最高水平,能提供少年、青年、老年等不同年龄段男女音库,支持普通话、台湾普通话和部分方言,并支持中英文混读。本文采用免费的 Microsoft Speech SDK 5.1 及其扩展包实现英文和中文的语音合成,它提供四种英文发音模式和一种简体中文发音模式。

5 面部表情和肢体运动控制

5.1 用户替身模型

用户替身模型需考虑面部和肢体结构,前者用于合成表情图像,后者用于合成肢体动作。3D 用户替身采用刚性连杆结构的虚拟人简化模型,如图 3 所示。机器人模型由髋关节、躯干、颈关节、颈部、头部、左右肩关节、左右上臂、左右肘关节、左右前臂、左右腕关节、左右手、左右大腿、左右膝关节、左右小腿、左右踝关节和左右脚组成。头、面部采用由规则曲面组成的简化结构。头部有双眼和口,用椭球或椭球冠模拟眼睑、眼球和口。眼睛有上下眼睑和眼球,上眼睑有一个沿水平中心轴的旋转自由度模拟闭眼动作;眼球有一个绕竖直中心轴旋转的自由度模拟视角的变化。上下嘴唇沿竖直方向有一个平移自由度,通过动态改变椭球冠短轴直径模拟口部动作。

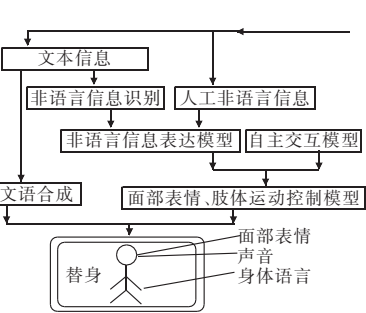


图 2 用户终端界面结构及可视语音合成过程示意

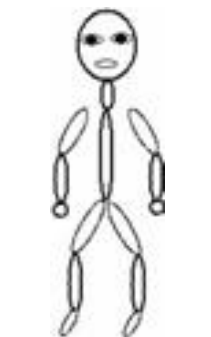


图 3 虚拟机器人模型

5.2 虚拟人肢体运动控制模型

虚拟人的肢体运动和面部表情控制采用程序驱动方法。连杆的运动具有继承关系。连杆 i 的运动由其父关节 $i-1$ 的绝对运动和连杆 i 绕该父关节的旋转相对运动合成。以矩阵 W_i 描述连杆 i 的运动,则有

$$W_i = W_{i-1} W_{Ri} \tag{1}$$

其中: W_{Ri} 为连杆 i 绕其父关节 $i-1$ 的旋转运动描述矩阵; W_{i-1} 为其父关节 $i-1$ 的绝对运动描述矩阵。以髋关节为根关节,记为关节 0。由递推关系,可得连杆 i 的运动描述矩阵为

$$W_i = W_0 W_{R0} W_{T1} W_{R1} \cdots W_{Ti-1} W_{Ri} \tag{2}$$

其中: W_0 为髋关节的运动描述矩阵; W_{Ti-1} 为关节 i 相对于关节 $i-1$ 沿连杆 i 平移一个连杆长度的矩阵描述。

虚拟人模型从髋关节到头部、左右手、左右脚共形成五条支链。由式(2)可得各支链中各个连杆的运动描述。

图 3 所示连杆结构虚拟人肢体运动对不同情感、意图表达的差异表现为各相邻连杆间沿自由度方向夹角及其角度变化

率的不同,因此通过配置各相邻连杆间夹角及角度变化率,能够实现各种非语言信息的肢体运动表达。以点头动作为例,其运动可视头部绕颈部沿水平轴向身体正前方旋转一个角度,再复位至零并重复 23 个周期。可以采用

$$\theta = A | \sin(\omega t) | \tag{3}$$

来控制,角度按照正弦规律变化。式(3)中: A 表征点头的幅度; ω 表征点头的速度。由于采用规则曲面构成头、面部,面部运动合成相对简单,仅包括眼睑、眼球和上下嘴唇的运动控制。上眼睑由椭球冠构成,能够绕其水平中心轴作正反向旋转,模拟眨眼动作。眼睑的开合也按照式(3)所示的正弦规律对旋转角度和速率进行调节,但要重新设置 A 和 ω ,使 θ 在时序空间上表现为一系列正弦脉冲。眼球由椭球组成,能够绕其竖直中心轴旋转,模拟虚拟人视角的变化;口部由椭球冠组成,在竖直方向上通过动态控制椭球短轴长度模拟上下嘴唇运动。

6 系统原型

采用 VB 6.0 和 OpenGL 实现的通信平台原型界面如图 4 所示。



图 4 基于可视语音合成的 3D 通信平台原型

该平台能够实现本机与本机或局域网内两台计算机之间的文本通信。进入系统后,输入要连接的计算机 IP 地址、设置本地端口和远程端口便可以与远程计算机建立连接。用户可在程序界面左下方文本框中输入信息并发送给远程计算机,接收到的文字在界面左上方文本框中显示。系统将接收到的文字合成为语音,同时虚拟场景中的机器人产生动作。在界面右下方的下拉框中,能够选择四种英文发音模式和一种简体中文发音模式。对接收到的简短文字信息,系统能够自动识别其蕴涵的非语言信息类别,并以适当的肢体动作表达。如接收到“你好”,虚拟人将鞠躬表达问候;接收到“是”“是的”等信息,虚拟人点头以示同意或肯定。

7 结束语

基于文语合成和可视语音技术,提出一种新型的即时文本通信方案。通信双方通过键入和发送文字交换信息,用户终端将接收到的文字合成为语音,同时接收发送端指定或在本地自动识别人类情感、意图等非语言信息类别,与相应的非语言信息表达模型进行匹配,通过 3D 用户替身的面部表情和肢体动作形象表示文字所蕴涵的人类情感和意图。建立了简短文字的非语言信息知识库,给出了非语言信息表达模型,给出了基于时间序列特征参数的用户替身自主交互模型。采用连杆结构的虚拟人肢体模型和由规则曲面构成的头、面部模型,采用程序驱动方法,以关节夹角和角度变化率为 (下转第 4214 页)

模块、基于特征多分辨率建模模块和建模核心三大模块组成。

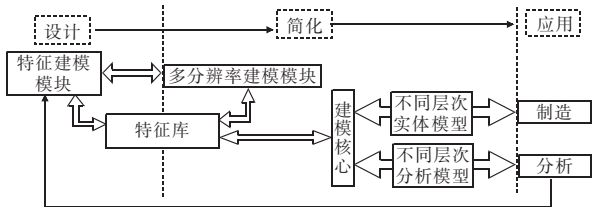


图 3 CAD-CAE 集成系统体系结构图

基于特征的建模模块主要用于模型的详细设计及提供特征操作接口。它将零件设计信息和工程信息综合处理建立起零件特征模型,并实现生命周期中形式特征的数据库管理。这部分可以采用任一主流商业特征造型系统或自主开发的一个特征造型系统。本文采用哈尔滨理工大学 CAD 研究所自主开发的造型系统 HUST-CAID 为造型平台。

多分辨率建模模块主要功能是管理多分辨率特征。这些多分辨率特征是用来实现简化过程,完成抽象分析模型管理。当用户定义一个 LOD 级别时,采用多分辨率造型算法提取多分辨率特征信息。

建模技术执行细节移除和降维处理等操作,从特征库中提取由基于特征的建模模块创建的模型数据,从中抽取相应级别的实体模型,也可以抽取一系列不同 LOA 层次的抽象模型。完成功能是创建和操作所有用于设计和分析的几何模型,以进行相应的设计和分析。其中 LOD 模型层次级由用户自定义。

3.2 系统建模模型中的并行机制

建模系统中的并行机制如图 4 所示。

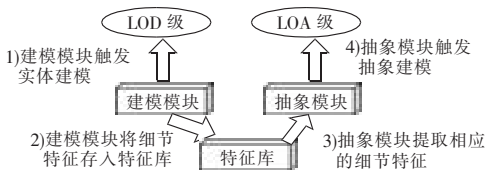


图 4 建模系统中的并行机制

用户首先通过基于特征的建模模块进行零件设计,特征建模操作将设计和分析所需要的特征几何模型合并到零件模型中,存储于特征库中。然后执行基于特征的多分辨率建模模块,将各 LOD 层次信息模型存储于特征库中。当用户定义一个 LOD_i 级后,建模模块将自动从特征库中取出模型,从中抽取相应的 LOD_i 实体模型和对应于 LOD_i 级的 LOA_j 级的抽象模型。

最后用户根据应用,对实体模型进行制造或渲染等操作,对抽象模型通过自动网格产生网格模型,并将它转换到 CAE 系统中进行分析,对设计进行优化。

4 结束语

本文提出了一个新方法来优化对于交互特征模型操作的几何约束求解。展示的约束求解驱动模型编辑方法不仅减少了用在交互求解中的约束数量,而且通过使已经存在的约束求解代替分析,避免了需要对分离约束模型分析的算法。事实上可以使用任何的求解方法,它具有求解一个欠约束模型和返回模型中是良约束的部分能力,这个特性在许多同时代的几何约束求解中都有。

总之,现在的方法在交互操作中优化了解几何约束,它可以被用于许多的约束求解器中,并避免了为分离约束模型分析所需的算法。原型执行描述和具有它的执行测试证明了此方法的潜力。

参考文献:

- [1] LEE S H. Feature-based multiresolution modeling of solids [J]. *ACM Trans on Graphics*, 2005, 24(4): 1417-1441.
- [2] LEE J Y, LEE J H, KIM H, et al. A cellular topology-based approach to generating progressive solid models from feature-centric models [J]. *Computer-Aided Design*, 2004, 36(3): 217-229.
- [3] LEE S H. A CAD-CAE integration approach using feature-based multi-resolution and multi-abstraction modeling techniques [J]. *Computer-Aided Design*, 2005, 37(9): 941-955.
- [4] DU Shi-hong, QIN Qi-min, WANG Qiao, et al. Evaluating structural and topological consistency of complex regions with broad boundaries in multi-resolution spatial databases [J]. *Information Sciences; an International Journal*, 2008, 178(1): 52-68.
- [5] KOO S, LEE K. Wrap-around operation to make multi-resolution model of part and assembly [J]. *Computers and Graphics*, 2002, 26(5): 687-700.
- [6] HOOKER J N. Logic, optimization and constraint programming [J]. *Journal on Computing*, 2002, 14(4): 295-321.
- [7] GAO Xiao-shan, JIANG Kun, ZHU Chang-zhu. Geometric constraint solving with conics and linkages [J]. *Computer-Aided Design*, 2002, 34(6): 421-433.
- [8] KIM D, KIM D S, SUGIHARA K. Apollonius tenth problem via radius adjustment and Mobius transformations [J]. *Computer-Aided Design*, 2006, 38(1): 14-21.

(上接第 4211 页)特征参数实现肢体运动控制与面部动作控制。采用 Microsoft text-to-speech SDK 及其扩展包实现文语合成。利用 VB 6.0 和 OpenGL 开发了具有可视语音合成功能的 3D 通信平台原型。

未来的研究包括非语言信息的智能识别、用户替身自主交互模型的完善、用户肢体运动和面部表情控制算法的改进、双/多用户虚拟场景研究及用户替身间交互行为模型等。

基于可视语音合成的文本通信技术综合利用了文本存储容量小、占用带宽少、传输速度快的优点;并利用语音合成技术实现语音增强,利用虚拟现实技术实现视觉增强,对接收到的文字进行形象的人物化朗读。基于可视语音合成的文本通信技术将成为新型通信平台的一个发展方向。

参考文献:

- [1] KUROSE H, ITO K, NISHIDA S. Study on expressing emotions for

reading texts with synthesized facial expression [C]//Proc of International Conference on Instrumentation, Control and Information Technology. 2005.

- [2] 王志明,蔡莲红,艾海舟.基于数据驱动方法的汉语文本-可视语音合成[J]. *软件学报*, 2005, 16(6): 1054-1063.
- [3] 尹宝才,李敬华,贾秉滨,等.基于两层隐马尔可夫模型的可视语音合成[J]. *北京工业大学学报*, 2006, 32(5): 416-418.
- [4] 杨志晓,徐朝辉,张德贤.基于虚拟和物理化身的中文文本信息具体化[J]. *系统仿真学报*, 2007, 19(10): 2287-2292.
- [5] YANG Zhi-xiao, GOFUKU A, ITO K. Text to scene: a concept of constructing expressive interface for text embodiment [C]//Proc of International Conference on Instrumentation, Control and Information Technology. 2005.
- [6] ZHE Xu, BOUCOUVALAS A C. Text-to-motion engine for real time Internet communications [C]//Proc of International Symposium on Communication Systems, Networks and DSPs. 2002: 164-168.