文章编号：2095-2163(2019)03-0232-05 中图分类号：TP18 文献标志码：A

基于 VR 的数字博物馆中语音交互设计探究

刘柏亨，原松梅
（哈尔滨工业大学 机电工程学院，哈尔滨 150001）

摘 要：随着 VR 技术的进一步普及，其应用场景也从娱乐扩展到医疗、文化等各个领域，数字博物馆便是在 VR 技术支持下应运而生的，具有文化传播价值。为了体现 VR 数字博物馆的交互性和代入感，语音交互设计一直是数字博物馆设计中急需解决的重要问题。本文从语音交互设计入手，对数字博物馆中如何实现语音交互设计以及用户体验进行探讨，以期为开发高水平的 VR 数字博物馆提供有益借鉴。

关键词：数字博物馆；交互设计；语音识别；VR

A research on voice interaction design in digital museum based on VR

LIU Baiheng, YUAN Songmei
(School of Mechatronics Engineering, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 With the further popularization of VR technology, its application scenarios have expanded from entertainment to medical treatment, culture and other fields. Digital museum emerged with the support of VR technology, and has the value of cultural communication. In order to reflect the interactivity and substitution of VR digital museum, the design of voice interaction is always an important problem that needs to be solved urgently in the design of digital museum. This paper discusses how to realize the voice interaction design and user experience in the digital museum, so as to provide some experience for the development of the high level VR digital museum.

【Key words】 digital museum; interaction design; speech recognition; Virtual Reality

0 引 言

时下，随着数字化技术的迅速发展，即使得基于移动终端的数字博物馆应运而生，真正突破了一时一地的时空局限，满足了人们足不出户、在手机终端浏览各地博物馆相关藏品信息的客观需求。但美中不足的是，数字博物馆仅仅是以文字、图片或视频的形式来呈现展馆内容，导致其体验感和沉浸感完全不及游览实体博物馆。

得益于虚拟现实（Virtual, Reality, VR）技术的迅猛发展，基于 VR 技术的数字博物馆不但突破了实体博物馆的时空局限，而且能以多感官、多层次和立体化的方式呈现展馆内容，弥补了其它终端临场感和代入感不强的缺陷；此外，VR 的交互性对用户具有更强的吸引力。

考虑到目前的场地和成本限制，现有的消费级 VR 交互方式仍是以手柄为主，在交互体验的自然性上表现较差，而作为自然交互方式的代表，语音交互一直都是 VR 交互研究的重要问题，且在 VR 数字博物馆中也具有较高的实用价值。据此，本文将针对语音交互设计在 VR 数字博物馆中的应用和用户体验进行探讨。对此可做研究论述如下。

1 VR 及其交互设计概述

1.1 VR 技术

VR 技术是一种计算机仿真技术，通过对三维世界的模拟创造出一种崭新的交互系统。其特点是能让用户以主角的身份进入到一种由计算机图形技术构成的、具有感知的虚拟空间环境中，用户通过借助 VR 设备与虚拟环境中的对象进行交互，以接近现实亲历场景的效果，对三维虚拟空间环境进行更真实的体验。

1.2 VR 中的交互设计

与在图形用户界面占据主流地位的视窗-图标-菜单-指针（Window-Icon-Menu-Pointer, WIMP）界面范式不同，VR 所遵循的是基于 VR 的交互（Reality-based Interaction, RBI），这一框架在 2006 年由 ACM CHI 会议的发起者 Jacob 等人[1]提出，主要包括物理学原理、人体感知与技能、环境感知与技能、社会感知与技能四个层次。

作者简介：刘柏亨（1995-），男，硕士研究生，主要研究方向：虚拟现实应用；原松梅（1960-），女，硕士，副教授，硕士生导师，主要研究方向：数字媒体技术、工程美学与数字仿真、大学计算机教育。

收稿日期：2019-01-18

从 Post-WIMP 到 RBI 范式的过程中，再没有出现类似 WIMP 一样稳居业界榜首的范式[2]，这是因为 VR 交互设备所使用的不再是单一、固定的离散型信息输入，而是呈现出多通道的特性，人们通过听觉、触觉甚至是味觉的感知，能够获得数倍于以往终端的信息量和丰富体验。

合适的输入设备对于 VR 的交互也同样重要，目前的离散输入设备、连续输入设备大多包含键盘、三维鼠标、力反馈手套、深度相机等种类，例如 Kinect 和 Leap Motion 等，而脑电波输入设备、语音和生理信号感知设备还不成熟，距离走出实验室尚需时日。

在消费级 VR 交互设备上，诸如按键手柄、深度相机等较为常见，其它交互设备由于连接复杂、不便携带等诸多限制仍然难以进入消费级市场，因此除了对视觉上的交互设计之外，对语音输入方面的交互设计研究也同样是现阶段的研究重点。

2　探究 VR 数字博物馆交互设计的必要性

2.1　数字博物馆建设的必要性

维基百科中给出的数字博物馆的定义为："数字博物馆是以博物馆为主题，结合多媒体技术应用的展示平台"[3]。随着人们生活水平的不断提高，文化消费成为了时下重要的消费内容之一，而参观游览博物馆就是文化消费的一种。同时，人们的生活方式发生了极大的变化，即使足不出户也可便捷享受高品质的现代服务及带来的优良体验，而更多喜爱艺术、文化和历史的人则有了在家中观看博物馆、美术馆等世界各地展馆藏品的需求，如此就使得数字博物馆的建设获得了发展契机。目前，科技的飞速进步证明：VR 技术作为数字博物馆设计中核心关键的技术，真正提供了突破时间和空间的限制、在虚拟世界中实现交互体验的可能。

2.2　VR 数字博物馆交互设计的必要性

近年来，各地博物馆在虚拟数字展馆建设上均有可观进展，虽然大部分博物馆囿于资金、人力资源等实际条件仍处在传统网站阶段，但已有博物馆开始着手或陆续加大了基于 VR 技术和相关平台的数字博物馆的研发投入力度，而且正处于快速发展的黄金阶段。

2017 年 10 月底，完全虚拟存在于 VR 的博物馆——克莱默博物馆（Kremer Museum）诞生了，并于 2018 年 3 月进入 HTC VIVE 的官方应用商城 VIVEPORT，人民币售价 37 元。作为一个极具创新意义的博物馆，克莱默博物馆将 VR 技术与世界级大师的绘画作品相结合，其中展示了荷兰黄金时代的许多泰斗级大师伦勃朗、克伊普、艾尔波特·盖依普和弗兰斯·哈尔斯的作品等，这也是世界上第一个完全在 VR 世界建立的博物馆，是 VR 技术在数字博物馆设计中的成功标志性应用。

与实际的博物馆场馆和传统网站阶段的虚拟博物馆不同，VR 技术应用的本身就意味着其交互的复杂性、多重性和更多的可能性。由于 VR 研究仍处于技术的更新演变期，其在虚拟博物馆上的应用也不仅仅只立足于展品和实体场馆场地的还原，VR 平台载体的创新、尤其是交互上的创新将为应用的内容本身带来更多的选择和设计空间。

2.3　VR 数字博物馆场景下语音交互研究的必要性

由于六自由度平台和触感手套等交互装备和相应交互方式受场地和购买成本等的限制，因此难以大规模进入消费级市场，而语音交互所需的设备门槛对于标准的 VR 设备而言并不高，目前的 VR 头显基本都配备了麦克风设备，不具备麦克风语音输入设备的 PC 头显也可以通过 PC 上的语音输入接口进行输入。

2016 年过后，随着 VR 技术相关研究水平的不断提升，众多实体博物馆都在积极推进与 VR 场馆相关应用的开展与落地，这些应用所面向的用户很少能使用类似触感手套等高成本的交互方式，语音交互就成为了除手柄交互外的最佳选择。

3　语音交互在 VR 数字博物馆中的应用研究

语音交互属于自然语言理解领域，是人工智能领域的分支之一。一个成熟的智能语音交互系统应包含语音识别模块、自然语言理解模块、自然语言生成模块、语音合成模块和对话管理模块。将以前沿研究成果 Deep-FSMN 模型为基础，开放式语音合成平台为辅，重点阐述在 VR 数字博物馆应用场景下的语音交互设计过程，研究过程详见如下。

3.1　Deep-FSMN 模型在语音识别的应用

阿里达摩院于 2018 年 6 月公开了一种改进的前馈型序列记忆网络（Feed-forward Sequential Memory Network，FSMN）架构，即 Deep-FSMN（DFSMN），将其应用在大词汇量的连续语音识别场景中，相比于 BLSTM 模型在各方面均具有一定优势。研究可知，这是一种基于 FSMN 模型的声学模型。

该模型是在 cFSMN 结构的基础上，通过在相邻

的存储块之间引入跳过连接层。这些跳过连接层则可以实现信息流向不同的层。其结构如图 1 所示。

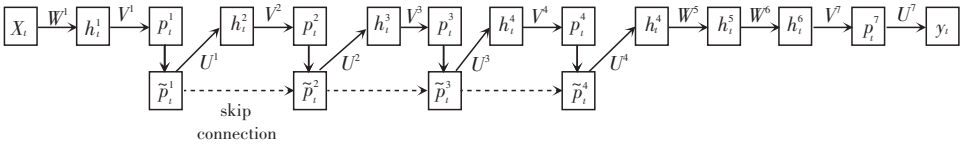

图 1 附加了跳过连接层的 Deep-FSMN 图[4]

Fig. 1 Illustration of Deep-FSMN (DFSMN) with skip connection[4]

由图 1 可见,在 cFSMN 层中,一个标准的隐层会被低秩权重矩阵分解为 2 个层,而一个 cFSMN 可解析为 4 个 cFSMN 层和 2 个 DNN 层,总共 12 个层,当需要通过增加存储块来进行高阶训练时,这种结构会导致梯度消失问题,故而特别引入了跳过连接层,这种设计旨在使低层梯度能通过存储块输出流到更高层的存储块。同时,在反向传播的过程中也能将高层的梯度分配到低层,而这将有助于提升识别效率。

这一模型现已开源至 github,支持通过搭建在线语音识别系统或从阿里云接入等方式来定制更高效的训练和语音识别功能,在 VR 交互中较多孤立词识别场景下可能会有更好的表现。

3.2 语音合成和 VR 空间声技术

虚拟环境应对用户的输入产生反馈,包括功能上的交互和语音应答等。在 VR 数字博物馆中,经常会遇到场馆解说词不便公开或难以采样处理的情况,现有的音频资料难以直接或经过处理后投付使用,因此采用语音合成(Text-to-speech,TTS)技术对文字资料进行转语音处理即是一种较为合理的解决方案。

与直接参考博物馆现有解说词和语音资料相比较,采用 TTS 技术有利有弊。即使借助目前的人工智能技术,在将 TTS 合成的语音与博物馆专业解说员的解说词进行对照后会发现,其在感情处理、语音语调(抑扬顿挫)上仍有很大差距。另外,专业的解说词经过了专门的写作润色和加工,是适合连续朗读的;而 TTS 的文字来源多样、且不一,会令用户产生不自然的感觉,进而带来与 VR 世界脱节的用户体验。

VR 环境与传统交互平台的一个鲜明区别就在于其交互对象的虚拟性,每一个交互都是发生在虚拟的三维空间中的,虚拟实体发出的声音需要模拟声音的空间位置和传播情况,因此,在选用双声道扬声器作为输出设备的前提下,应在开发环境中使用空间音频。

VR 空间声技术是在三维音频技术的基础上衍生而来,其中一个关键技术就是 VR 三维音频渲染技术将采集、解码得到的声道、对象和声场信号在 VR 设备上重放,达到真实感和空间感兼具的听觉体验。Ambisonics 音频文件经过解码之后再次还原成一个空间声场,此声音相当于是从球形空间中各个方位的虚拟扬声器(Virtual Speakers)上发出来的[5]。在本类系统的应用场景中,扬声器多为双声道立体声扬声器,大体上可分为 PC 端桌面扬声器和立体声耳机两种,VR 空间声技术的虚拟扬声器则恰好作为发声的虚拟实体的映射存在于 VR 博物馆的场景中。

3.3 基于 VR 的数字博物馆中语音交互设计研究

本研究拟以解放战争三大战役之一的辽沈战役纪念馆为目标场馆,通过搭建基于 VR 平台的数字博物馆实验,并将使用语音指令控制和语音交互来完成整个体验和游览过程。

基于前述研究成果,本实验将运行在 HTC VIVE 上,同时采用 Unity 2018 进行基础性的虚拟资产搭建[6],以及基本漫游功能的配置,通过与 HTC VIVE 自带手柄相结合的交互方式,实现在游览过程中的语音交互。

文中的 VR 博物馆的语音输入交互基本流程如图 2 所示。由图 2 可知,当用户需要操作控制器进行漫游时,即按下映射了脚本的手柄控制器,使语音识别系统进入语音激活检测(Voice Active Detection,VAD)状态,保持激活检测状态,输入语音信号,经过录音脚本传输至识别模块,识别后再将结果作为文本输出。交互逻辑则需要以快速迭代模式进行开发,首先梳理基本语音指令控制逻辑,继而将针对目标 VR 博物馆中的内容进行扩充。

考虑到针对 VR 应用场景的语音交互,尤其是语音指令控制功能的交互,将默认使用中文语料进行训练,本系统将满足辽沈战役纪念馆的游览和交互使用,语料中除基本的交互常用词和高频词之外,还有该馆的场馆名、主要和具有代表性的藏品名及

相关背景的重点名称等，这些信息在大部分公共的汉语普通话语料库中有所收录，文中对其阐释分析如下。

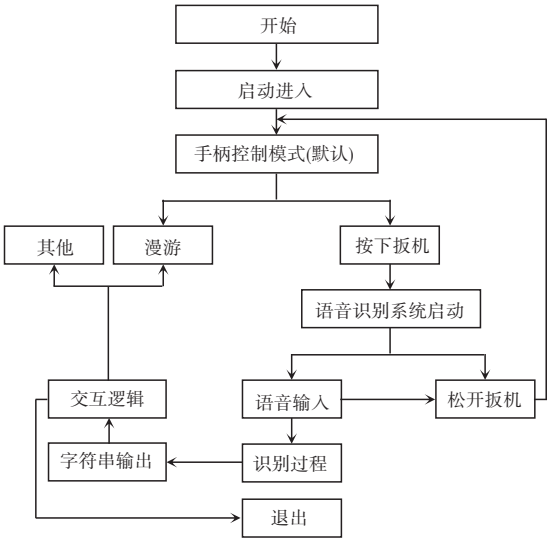

图 2　VR 博物馆的语音输入交互基本流程图

Fig. 2　Basic flow chart of voice input interaction of the VR museum

目标场馆、也就是辽沈战役纪念馆占地 18.8 万平方米，以辽沈战役军事主题为切入点，其中《攻克锦州》是中国第一座全景画馆，被誉为中国博物馆和世界美术史的艺术精品和经典之作[7]。这些重点城市和战役等内容，马云飞等重点人物、连同战史馆、支前馆等馆名及藏品名，如果作为单个孤立字、词加以识别，虽然在字本身的识别上可以取得最优结果，但对于该类专有词汇识别效率并不高。因此通过借鉴游戏中语音指令控制的研发经验，可授权开发者或管理员能够自行定制专门的语料，便于在同类博物馆之间的快速应用与移植。

从图 2 可以看出，本系统使用了与 VR 头盔设备相结合的方式进行语音交互，以 HTC VIVE 平台的交互设备为核心，期望实现手柄与语音指令同步交替控制的理想体验。

与 VR 技术相结合的语音交互所涉及的不仅是交互场景的改变，更是由实体的交互转向与虚拟实体的交互，是一个新的开始，对其交互体验的评价和研究也不能凭借单一结果或数据的衡定与考察，而应采用以量化评价标准为主、用户体验为辅的多样化研究方式来进行科学系统的综合评估。

4　交互体验和评价研究

4.1　语音识别的评价

对于这一类的在线语音识别，通过输出识别结果的字符串进行评价，将其作为单独的语音识别系统做出评价，同时，以纯识别系统的识别率作为主要参考标准，而在一般情况下，这一识别率的技术指标就是词错误率(Word Error Rate，*WER*)。

为了使识别出来的词序列和标准的词序列之间保持一致，需要进行替换、删除或者插入某些词，这些插入、替换或删除的词的总个数，除以标准的词序列中词的总个数的百分比，即为 *WER*，其数学公式可表示为：

$$WER = 100 \times \frac{S + D + I}{N}\%$$

其中，*S* 为 Substitution，即替换词个数；*D* 为 Deletion，即删除词个数；*I* 为 Insertion，即插入词个数；*N* 为总单词数。

但在实际使用中，语音识别的效率也至关重要。已有研究表明：在线环境或是 VR 场景中，语音的录制和传输都会产生延迟，而交互中的语音指令控制对交互反馈的时间要求较高，当采用了 Deep-FSMN 声学模型后，不仅在一定程度上提高了识别效率，而且也减少了建模过程的声音信号损失。因此，在进行评价时也需要反馈时间。

4.2　语音合成的评价

语音合成技术将文本转化为声音，广泛应用于多种场合中。其实现需用到语言学、语音学的诸多知识，不同的 TTS 系统在准确性、自然度、清晰度、还原度等方面也有着不一样的表现，因此，本系统的评价标准主要由 3 个部分组成，可对其分述如下。

(1)发音准确性。线上语料中频繁出现的多音字、数字、符号、夹杂英文等会给 TTS 带来挑战，具体如图 3 所示。这类情况的发音错误会导致较差用户体验。发音准确是确保用户体验的基本要求，现有的 TTS 系统已可以保证在交互场景下基本的发音准确性。

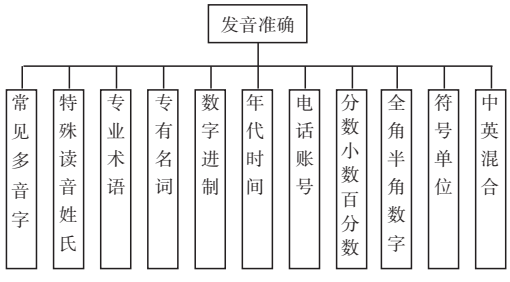

图 3　影响 TTS 发音准确性的因素

Fig. 3　Factors affecting the accuracy of TTS pronunciation

(2)韵律准确性。前端文本处理过程中会对文本进行分词处理和时长预估,为准确评估 TTS 停顿和发音时长的合理性,可以准备不同领域、不同句式、不同情感的文本,通过众测主观判断合成语音是否可接受,计算 TTS 韵律准确性。

(3)平均主观意见分(Mean Opinion Score, MOS)。业界对语音的整体评测一般使用 MOS 作为标准。在邀请听音人试听合成语音后,听音人根据分值描述,从拟人性、连贯性、韵律感等方面为语音选择合适的评判分数。

辽沈战役纪念馆承载着丰富的历史内涵,其解说词多具有较为充沛的情感,但由于目前技术原因,时下的 TTS 语音合成的拟人性和情感仍然属于大样本训练的结果,而非真正的人性化的情感,因此在拟人性的评测上应适当放宽要求。

4.3 VR 交互体验评估

这一部分采用问卷调研的方式,问卷设计基本思想遵循的是 VR 研究经典问卷(Presence Questionnaire, PQ)。

VR 研究经典问卷是由美国陆军研究所的 Witmer 等人于 1992 年提出,并于 1998 年再次更新、且通过了可靠性验证,由此将提升临场感的因素分为控制因素、感官因素、分神因素和真实度因素四类[8]。在 VR 交互上主要是从控制因素方面进行问卷设计。

通常而言,控制因素包括控制程度、控制的直接性、可预期性和控制模式。对于虚拟环境而言,用户对任务环境的控制越符合自然习惯,控制程度越强,给虚拟环境带来的变化越明显、也越容易预测,用户的临场感随即也就越强。遵循这一原则,本研究尝试在目标场馆的应用场景下进行用户体验评价问卷的设计。

VR 语音交互问卷问题分类见表 1。表 1 中给出了 5 个基本的问题分类,采用李克特量表对用户进行调研。其中,每个分类可拆分成多个细节问题,主要监测了该系统的功能性体验,例如 Q1 和 Q5 从一定程度上考察了语音识别的效率和表现情况,Q2 考察了三维音频的体验,Q3 则考察 TTS 功能,这些问题细化后即成为一份完整的问卷,从场馆内容的针对性、交互的可用性等方面进行全方位的研究,以期从每个问题中寻求对应的改进方案。

表 1 VR 语音交互问卷问题分类

Tab. 1 Question classification in VR voice interaction questionnaire

编号	问题大类	差 (1 分)	一般 (2 分)	尚可 (3 分)	满意 (4 分)	惊喜 (5 分)
Q1	能否接受语音发出后的响应延迟					
Q2	虚拟环境中声音音效的匹配程度					
Q3	解说词内容和表现情况					
Q4	手柄交互和语音交互的切换体验					
Q5	语音交互指令完成情况					
…	…					

5 结束语

基于 VR 的数字博物馆已成为当下 VR 应用领域的研究热点。本文即以 Deep-FSMN 模型为基础,通过应用场景的针对性训练、TTS 与实体馆语音素材的混合应用及量化标准和主观评测的结合,在 VR 博物馆场景下语音识别效率的提升上有一定进展,并实现了 VR 数字博物馆的语音交互体验的优化,为后续研究和开发更为成熟的 VR 数字博物馆提供了有益的支持。诚然,VR 技术及应用领域还有广阔的探索研发空间。值得期待的是,其在数字博物馆的开发设计及其文化传播中必将发挥更大的作用。

参考文献

[1] JACOB R J K, GIROUARD A, HIRSHFIELD L M, et al. Reality-based interaction: A framework for Post-WIMP interfaces[C]// Proceeding of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems (CHI '08). Florence, Italy: ACM, 2008: 201-210.

[2] 张凤军, 戴国忠, 彭晓兰. 虚拟现实的人机交互综述[J]. 中国科学: 信息科学, 2016, 46(12): 1711-1736.

[3] 维基百科. 数字博物馆[EB/OL]. https://zh.wikipedia.org/wiki/虚擬博物館.

[4] ZHANG Shiliang, LEI Ming, YAN Zhijie, et al. Deep-FSMN for large vocabulary continuous speech recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5869-5873.

[5] 胡瑞敏, 王晓晨, 张茂胜, 等. 三维音频技术综述[J]. 数据采集与处理, 2014, 29(5): 661-676.

[6] 丛晓丹, 吴冈, 管练武. 基于 Unity3D 的数字纪念馆虚拟漫游设计[J]. 自动化技术与应用, 2017, 36(11): 85-88, 92.

[7] 辽沈战役纪念馆. 辽沈战役纪念馆官方简介[EB/OL]. [2017]. http://www.jzlszy.com/index.php? m = page&a = index&id = 132.

[8] WITMER B G, SINGER M J. Measuring presence in virtual environments: A presence questionnaire [J]. Presence: Teleoperators and Virtual Environments, 1998, 7(3): 225-240.