基于语音识别控制的人体运动仿真

王志文[1]，张子豪[2]，贾玉祥[1]，魏毅[3]，沈燕飞[4]
(1. 郑州大学信息工程学院，河南 郑州 450001；2. 中国科学院计算技术研究所，北京 100190；
3. 武夷学院数学与计算机学院，武夷山 354300；4. 北京体育大学信息工程学院，北京 100084)

摘要：人体运动仿真是指在计算机中实时地渲染人体及其动作，是建立计算模型模拟人体在给定约束下自然真实的物理运动。渲染良好的人体运动，能够提高人们在观看时的沉浸感与代入感，同时，还是观者与系统之间交互的必然前提。基于上述理论，研究在使用渲染引擎，保证人体运动渲染质量的前提下，进一步提出*使用语音识别引擎对观看者的口令进行识别，增加观者与系统之间的交互*。经过仿真实验，结果证明系统可行。

关键词：语音识别；人体模型；人体运动仿真；人机交互
中图分类号：TP391 文献标识码：A 文章编号：1004-731X (2018) 11-4203-07
DOI: 10.16182/j.issn1004731x.joss.201811019

Human Motion Simulation Based on Speech Recognition Control

Wang Zhiwen[1], *Zhang Zihao*[2], *Jia Yuxiang*[1], *Wei Yi*[3], *Shen Yanfei*[4]
(1. School of Information Engineering Zhengzhou University, Zhengzhou 450001, China; 2. Institute of Computing Technology Chinese Academy of Sciences, Beijing 100190, China; 3. Wuyi University School of Mathematics and Computer Science, Wuyishan 354300, China; 4. School of Information Engineering, Beijing Sport University, Beijing 100084, China)

Abstract: Human body motion simulation refers to real time rendering of human body and its motion in a computer. It is a computational model to simulate the natural and physical movement of human body under given constraints. Rendering a good human body movement can improve people's sense of immersion and substitution when watching. It is an inevitable premise of interaction between the viewer and the system. Based on the above theory, under the premise of using the rendering engine to ensure the quality of human motion rendering, *it is further proposed to use the speech recognition engine to identify the viewer's password and increase the interaction between the viewer and the system.* The results of simulation experiments prove that the system is feasible.

Keywords: speech recognition; mannequin; human body motion simulation; human-computer interaction

引言

人体仿真运动应用到三维模型上时，能够高程度实现现实运动的模拟，当用户处于高仿真的动画场景中，代入感增强，容易沉浸于虚拟场景。通过人机交互系统，控制虚拟人物运动，通过增加语音控制方式，简化虚拟人物操作，减少现实中的肢体动作，减少虚拟现实的隔离感。

人体运动仿真是一种利用计算机模拟自然真实人体运动过程的技术[1]。具体包括建立计算模型，仿真计算虚拟人在给定约束条件下自然真实的物理运动过程，并在计算机生成的虚拟环境中以三维模型式逼真呈现该运动过程[2]。

收稿日期：2018-05-25 修回日期：2018-06-28；
基金项目：国家自然科学基金(61402419)；
作者简介：王志文(1995-)，男，安徽池州，研究方向为运动仿真；张子豪(1993-)，男，四川成都，博士生，研究方向为人体运动仿真和人脸动画等；贾玉祥(1981-)，男，河南沈丘，博士，讲师，研究方向为自然语言处理。

三维图形引擎在图形系统、虚拟现实和游戏中都被广泛运用[3]。一个好的三维图形引擎需要良好的动画模块的支持，在动画模块中，运用适当的技术和算法，使动画流畅并接近于实际情况。使用三维图形方式实现的人体运动仿真可应用于动画和影视特效、交互式游戏、生物力学等领域。

随着虚拟现实技术的发展，人们已经不再满足于作为尝尽的观看者。人们更愿意参与到所能观看到的场景当中，这不仅要求我们对场景、人体渲染的质量达到以假乱真的水平，还要求我们能够按照符合人体日常习惯的方式去安排人与场景之间的交互。而人们日常与场景最多的交互，除开眼睛接收场景信息以外，就是通过声音传达以及接收信息了。因此，基于语音的交互，是十分有必要的。

语音识别，以语音为研究对象，是语音信号处理的一个重要研究方向，是模式识别的一个分支，其研究目的就是要让机器具有人的听觉功能，在人机语音通讯中“听懂”人类口述的语言[4]。语音识别技术与其他自然语言处理技术(如机器翻译及语音合成技术)相结合，可以构建出更加复杂的应用。

当前大多数软件应用语音接口进行网络通信和语音文本转换[5]，作为交流方式的一种。基于语音接口的人机交互技术尚未成熟，因语音反馈的结果的多样性，人工智能产品常常很难做出正确的应对。其中，基于语音控制的三维动画变换也是处于发展之中，通过语音接口控制三维人物模型动画的转换，使得人机交互更加便捷，应用到游戏领域中，将简化玩家的手动操作。

现有的动画引擎不提供语音命令接口，且没有成熟的解决方案，只能通过接入第三方语音平台来实现语音控制接口。

基于以上技术基础，本研究提出了基于 Unity 引擎渲染的人体运动仿真引擎，并且这套引擎是高度可扩展的，同时集成语音识别技术，提出了基于语音交互控制的人体运动仿真的实现。

1 三维模型动画

1.1 模型驱动

常用模型驱动算法有编辑几何模型的线性混合蒙皮算法、复杂人体模型线性蒙皮算法、基于骨骼关节的自动绘制算法等。其中，LBS 算法要求 2 模型骨架需要手动绑定，应用到该设计的三维模型上时变形效果不理想[6]。

在传统 LBS 变形算法中，顶点的坐标位置由以下公式给出：

$$v' = \sum_{i=1}^{n} w_i C_{j_i} v$$

即，针对顶点 v 的变形由所有影响该顶点的关节点 C_{j_i} 以及对应权重给出 w_i，这种变形算法的缺陷在于在一些特定的情况下，面片会出现塌陷的情况[7]。因此基于模型变形的工作，引入了对偶四元数 $\hat{q} = \cos\frac{\theta_0}{2} + s_0 \sin\frac{\theta_0}{2} + \varepsilon(\mathrm{s}_\varepsilon \sin\frac{\theta_0}{2} - \frac{\theta_\varepsilon}{2}\sin\frac{\theta_0}{2} + \mathrm{s}_0 \frac{\theta_0}{2}\cos\frac{\theta_0}{2})$，其中 s_0 是旋转轴，θ_0 是旋转角，θ_ε 是沿旋转轴的平移，$s_\varepsilon = r \times s_0$，$r$ 是旋转中心[8]。算法描述如下：

塌陷解决算法描述

输入：对偶四元数 $\hat{q}_1, \cdots, \hat{q}_p$，顶点位置 v 和常数 v_n，关节指数 $j_1, \cdots, j_n$ 和权重 $w_1, \cdots, w_n$。

输出：修改后的顶点位置 v' 和常数 v'_n

$\hat{b} = w_1\hat{q}_{j_1} + ... + w_n\hat{q}_{j_n}$

//其中 $\hat{b}$ 的非偶数项为 b_0，偶数项为 b_ε

$c_0 = b_0 / \| b_0 \|$

$c_\varepsilon = b_\varepsilon / \| b_0 \|$

//c_0 标量表示为 a_0，c_0 向量表示为 d_0

// c_ε 标量表示为 a_ε，c_ε 向量表示为 d_ε

$\boldsymbol{v}' = \boldsymbol{v} + 2\boldsymbol{d}_0(\boldsymbol{d}_0 \times \boldsymbol{v} + \boldsymbol{a}_0\boldsymbol{v}) + 2(\boldsymbol{a}_0\boldsymbol{d}_\epsilon - \boldsymbol{a}_\epsilon\boldsymbol{d}_0 + \boldsymbol{d}_0 \times \boldsymbol{d}_\epsilon)$

$\boldsymbol{v}'_{\boldsymbol{n}} = \boldsymbol{v}_{\boldsymbol{n}} + 2\boldsymbol{d}_0 \times (\boldsymbol{d}_0 \times \boldsymbol{n}_{\boldsymbol{n}} + \boldsymbol{a}_0\boldsymbol{v}_{\boldsymbol{n}})$

//得到的 $\boldsymbol{v}'_{\boldsymbol{n}}$ 必须通过逆矩阵、转置矩阵进行变换

1.2 动画

三维动画大体分为关节动画、关键帧动画、骨骼动画三种。关节动画即把角色模型进行分割，独立成若干部分，每一个部分对应一个网格，独立部分的动画通过连接，生成一个整体的动画，产生的角色比较灵活；关键帧动画是由一个完整的网格模型构成，在整个模型动画序列的关键帧里记录网格的各个顶点的原位置及其改变量，然后进行插值运算，实现三维角色动画效果，生成的角色动画较真实；骨骼动画是应用最广泛的动画方式，它集成了关节动画和关键帧动画的优点，骨骼根据角色的特点组成一定的层次结构，使用关节连接各个骨骼，每一节骨骼可做相对运动，而皮肤作为单一网格蒙在骨骼外层，体现为角色的外观[9]。

本设计中使用的 Unity3d 引擎，由于其渲染机制原因，通过修改三维模型拓扑结构生成三维模型动画的方法会导致三维模型变形[10]。因此，在生成动画的时候使用 LBS 算法计算三维模型变形。

使用 Unity 中的动画系统，对人体运动动画进行剪切，生成多个不同的基本动作动画片段，如图 1 所示。

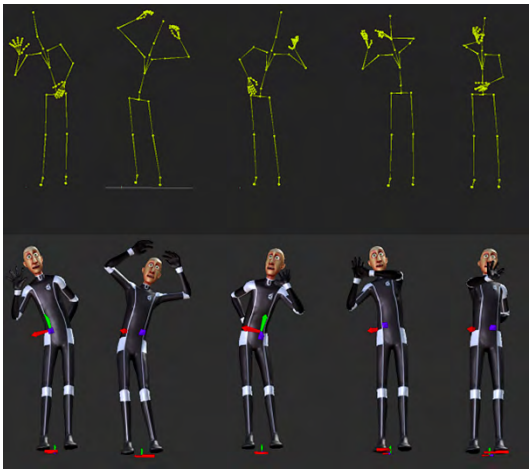

图 1　骨骼驱动的模型动画
Fig. 1　Skeleton-driven model animation

对于不同运动动画切换的平滑处理，需要在 Unity 动画状态机中进行设置相关参数值[11]。一般创建好 animation controller 之后，Unity 中对于动画的过渡会自动处理，如果过渡方法不能满足需求，需要开发者在状态机中自定义动画转换的时间、动画转换时的目标动画播放帧、过渡时间以及过渡时的优先级等参数，如此便可通过 Unity 自带的动画系统实现动画切换时的平滑处理。

2　技术集成

整个系统以三维动画引擎集成语音识别技术为核心，使用三维动画引擎生成骨骼运动数据驱动的人体模型动画，实现基于语音接口控制的人体仿真动画。

2.1 语音识别引擎

语音识别主要由端点检测、特征提取、模式匹配等几个部分组成[12]。端点检测就是要准确的检测出语音信号的起始点，将语音信号与噪声信号区分开。特征提取即寻找能够有效表示语音特征的参数模式匹配(即计算一段语音与特征模板之间的相似度)。经过对原始语音信号的流程化处理，得到最终的识别结果。整体流程见图 2。

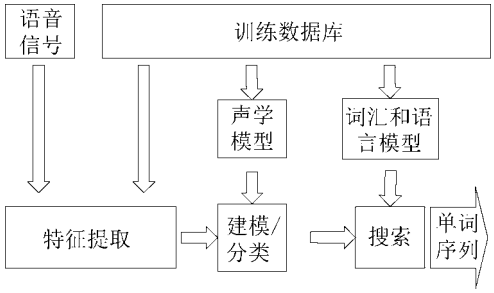

图 2　语音识别引擎流程图
Fig. 2　Speech recognition engine flow chart

近几年将机器学习领域深度学习研究引入到语音识别声学模型训练，使用带 RBM 预训练的多层神经网络，极大提高了声学模型的准确率[13]。目前国内提供语音识别服务的平台都采用了最新的语音识别技术，通过产品向用户提供调用接口。

2.2 动画系统与语音接口的集成

动画系统和百度语音接口的集成重点在于 Unity3d 中集成语音模块[14]。其中，百度语音提供

了完整的 SDK 开发接口。

语音接口集成到 Unity 的两种方案：

(1) 依据百度语音官网所提供的 REST API；

(2) 使用官网提供的 Android SDK。

2.2.1 REST API

REST API，支持普通话、粤语、英文 3 种语种，需要上传完整的录音文件，且录音文件时长不超过 60 s，支持上传的文件格式有 pcm (不压缩)、wav (不压缩，pcm 编码)、amr (压缩格式)三种。

在 unity 的脚本中通过 REST API 调用语音识别接口之前，需要根据鉴权认证机制获取认证。获取正确的令牌后，录制语音，将语音转换成 base 64 编码的字节流数据[15]。然后，将语音格式、采样率、声道数、token 令牌等参数进行 json 数据格式的包装，通过 POST 请求进行上传，得到反馈的识别结果。此外，还可以将录音数据放到 http body 中，定义请求头数据类型，再进行带请求头的网址访问获取识别结果。两种数据上传方式获得的结果无差别，通过 json 分析便可提取出语音识别结果的文本[16]。Unity 和语音服务交互过程如图 3 所示。

2.2.2 Android SDK

Unity3D 与 Android SDK 的集成需要额外 Android 工程开发的知识提供支持，整个集成过程如图 4 所示。

Android SDK，采用流式协议对语音数据进行处理，使用边处理边反馈的方法解析语音。相对于 REST API 全平台支持，Android SDK 只支持 android 平台。

Android SDK 无法直接调用，开发者需要建立 Android 库工程进行自定义类的编写，调用 SDK 的事件类，以此来实现对语音识别功能接口的调用[17]。自定义 Android 库完成后，通过 Unity 与 Android 之间的通信机制将语音识别技术集成到引擎中[18]。

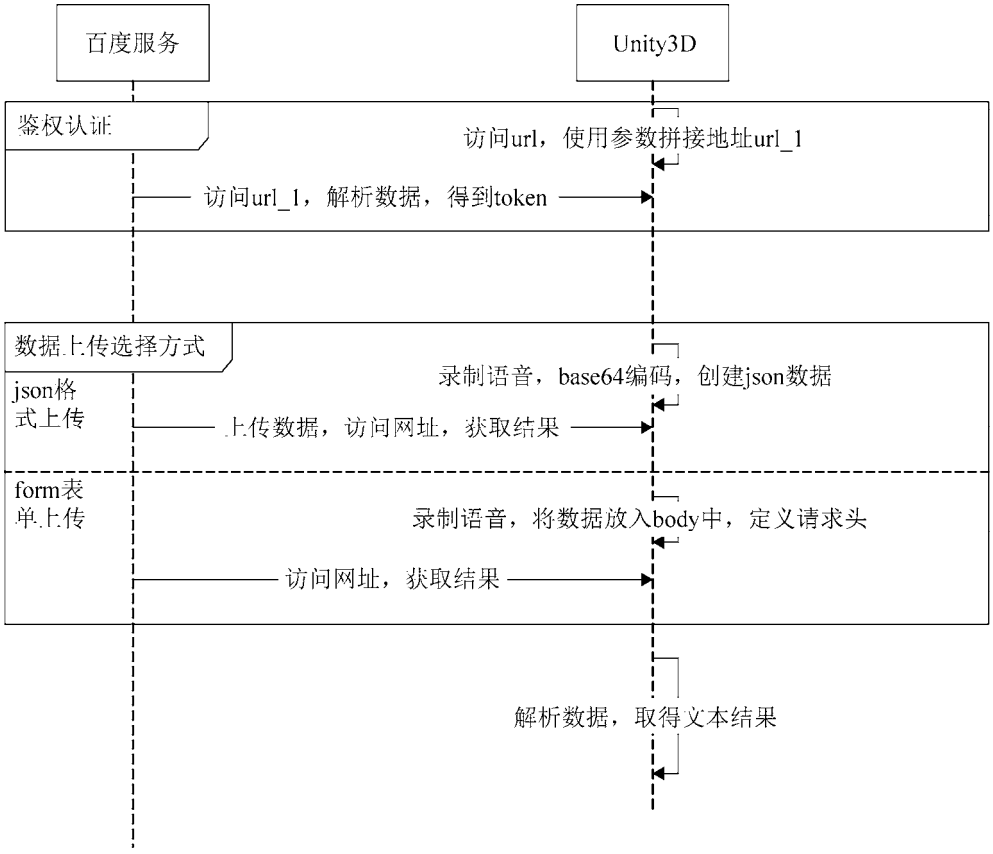

图 3 语音服务与 unity 交互

Fig. 3 Voice service interaction with unity

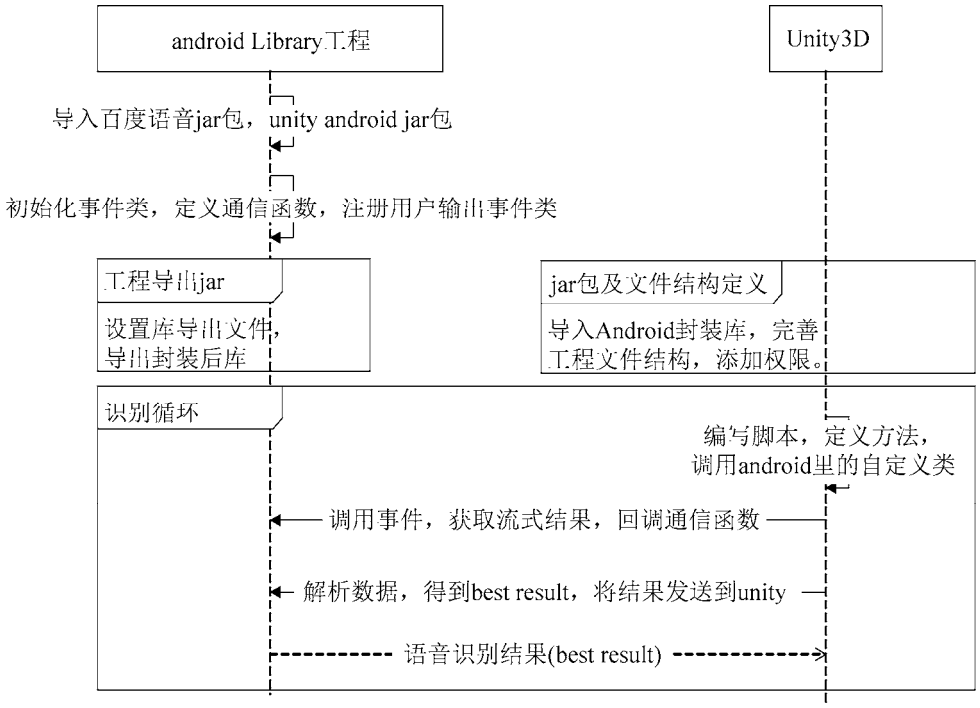

图 4　Unity 集成 Android SDK
Fig. 4　Unity integrated Android SDK

2.3 语音控制转换

编写脚本文件, 调用实现好的百度语音识别方法, 获取语音识别反馈的文本结果, 分析出相应的指令, 通过判断口令来改变动画 BTree 的条件变量, 实现动画的过渡及转变, 达到语音输入控制人物模型进行仿真运动的目标[19]。

动画和语音集成模块完成之后, 在 unity 开发平台上进行模块功能的整合, 实现语音控制三维模型进行仿真动画的转换[20]。

在 Unity 项目中导入已绑定人体骨骼的三维模型, 使用骨骼动画系统, 使用骨骼运动数据帧驱动三维模型, 添加相关组件, 完成人物模型动画的创建。在 Unity3D 中创建简单的场景, 添加语音识别技术模块的代码, 获取调用语音识别技术后的反馈结果, 与动画控制口令进行比对, 实现三维模型动画转换, 以达到语音控制动画转换的效果[21]。

3 实验及分析

实验内容主要分为 3 个部分, 一是用户试用评价, 二是语音识别两种方式集成到 unity 之后的结果反馈时间测试, 三是语音控制 Unity 场景人物模型动画的测试。

3.1 语音接口时间响应测试

REST API 和 Android SDK 集成到 Unity 之后的测试结果如图 5 所示。(时间为估计值, 误差 ±0.5 ms)

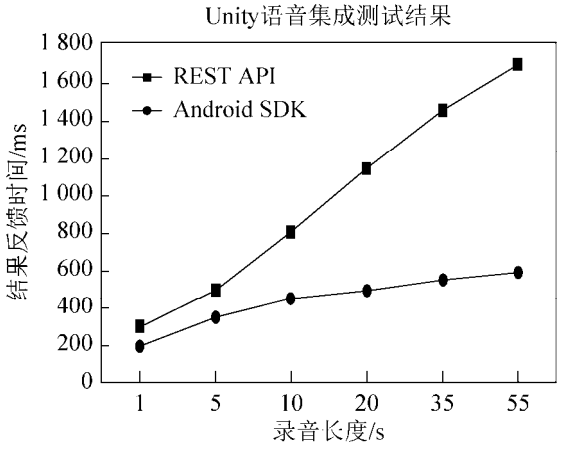

图 5　语音识别集成后的测试结果
Fig. 5　Speech recognition integrated test results

产生这样的结果主要是因为 REST API 需要提供完整录音才开始识别, 而 Android SDK 则采用流式识别，录音的同时就在开始识别，实时反馈识别结果。REST API 只返回一个 json 格式数据，Android SDK 返回多个 json 数据, 但最终会有一个 best result。

就反馈速度来说，Android SDK 快于 REST API，但就支持平台来说，REST API 支持全平台开发。因而，这里对于语音识别 SDK 的选择取决于 unity 发布平台，若发布到 android 系统上则选择 Android SDK，若为其他平台(IOS 除外)则选择 REST API。

3.2 语音控制动画切换测试

在 unity 中建立场景，将生成的仿真动画剪切成走路、跑步、跳、招手、OK 五个动画片段，添加语音输入接口和触屏输入接口，如图 6 所示。

图 6 系统界面和自定义模型姿态展示
Fig. 6 System interface and model gesture display

编辑动画状态机，添加变量条件，通过录入语音口令“跑”、“跑步”、“跳”、“走”、“走路”、“招手”、“OK”，测试不同动画转换反馈时间，观察不同动画之间转换的流畅性，测试结果如表 2 所示。

表 2 Unity 语音控制动画转换应答时间
Tab. 2 Voice control animation response time

人物动画类型	系统应答时间/ms
跑(跑步)	600
走(走路)	500
跳	400
招手	650
OK	450

经过测试, 人物模型在语音控制下完整完成了动画变换，且动画效果良好，转换无问题。

3.3 用户评价

本设计完成后，根据可交互性、真实感以及整体评估 3 个维度进行用户试验。向用户提供有语音和无语音驱动的人体运动仿真系统，对比 3 个维度，对两种系统进行分数评定以及使用评价。

参与本次用户评价的人数为 121 人，主要集中于在职人员, 其中专业领域 41 人, 普通群众 80 人, 评分总分为 100，评价结果如表 3 所示。(数据结果保留小数点后一位。)

表 3 用户评价平均分结果
Tab. 3 User rating average score results

维度	语音驱动	无语音驱动
可交互性	87.5	63.7
真实感	90.4	89.5
整体评价	88.4	70.6

从表 2 的平均分结果可知, 有语音驱动的系统在可交互性和整体评价中分数高于无语音驱动的系统得分, 两个系统的仿真动画得分很高且差别不大。由此说明有语音的人体仿真运动的确提高了人机交互的便捷性，而且不影响模型动画的仿真程度, 进一步说明了本研究应用是有效的，得到的结果符合预期。

4 结论

本文将通过深度摄像头采集的人体运动数据应用到 unity3d 的虚拟人物模型上，集成百度语音识别第三方语音服务平台，实现了语音控制虚拟人物三维仿真动画的转换。实验表明，通过语音作为输入接口，控制虚拟人物仿真运动的方法是有效的。

虽然文中只使用了百度平台的语音服务, 但基于其他平台的语音服务，在技术上集成到 Unity3D 是可行的。此外，Unity3D 中骨骼动画系统使用现有的骨骼数据驱动模型动画, 在一定程度上提高了仿真程度, 对于 Unity3D 使用骨骼数据生成骨骼动画的技术点还仍有提升的空间。相信在不久之后，

语音服务和仿真动画的集成开发更加成熟, 应用领域也将更为广泛。

参考文献:

[1] 夏时洪, 魏毅, 王兆其. 人体运动仿真综述[J]. 计算机研究与发展, 2010, 47(8): 1354-1361.
Xia Shihong, Wei Yi, Wang Zhaoqi. Summary of human motion simulation[J]. Computer research and development, 2010, 47(8): 1354-1361.

[2] Xia S, Gao L, Lai Y K, et al. A Survey on Human Performance Capture and Animation[J]. Journal of Computer Science and Technology (S1000-9000), 2017, 32(3): 536-554.

[3] 李胜亮. 三维图形引擎关键技术研究[D]. 西安: 西北工业大学, 2007.
Li Shengliang. Research on Key Technologies of 3D Graphics Engine [D]. Xi'an: Northwestern Polytechnical University, 2007.

[4] 李晓霞, 王东木, 李雪耀. 语音识别技术评述[J]. 计算机应用研究, 1999(10): 1-3.
Li Xiaoxia, Wang Dongmu, Li Xueyao. Speech recognition technology review [J]. Application Research of Computers, 1999(10): 1-3.

[5] 黄永峰. 因特网语音通信技术及其应用[M]. 北京: 人民邮电出版社, 2002.
Huang Yongfeng. Internet voice communication technology and its application [M]. Beijing: People Post Press, 2002.

[6] Loper M, Mahmood N, Romero J, et al. SMPL: a skinned multi-person linear model[C]// ACM SIGGRAPH Asia Conference. ACM, 2015: 248.

[7] Kavan L, Collins S, Žára J, et al. Geometric skinning with approximate dual quaternion blending[J]. ACM Transactions on Graphics (S0730-0301), 2008, 27(4): 105.

[8] Jacobson A, Baran I, Sorkine O. Bounded biharmonic weights for real-time deformation[C]// Acm Siggraph. ACM, 2011: 1-8.

[9] 王茂松. 三维引擎动画关键技术的研究和实现[D]. 武汉: 华中科技大学, 2014.
Wang Maosong. Research and Implementation of Key Technologies of 3D Engine Animation [D]. Wuhan: Huazhong University of Science and Technology, 2014.

[10] Baran I. Automatic rigging and animation of 3D characters[C]// ACM SIGGRAPH. ACM, 2007: 72.

[11] 张美香, 郝轶鸣. 关键帧动画技术综述[J]. 山西广播电视大学学报, 2009, 14(5): 55-56.
Zhang Meixiang, Hao Yiming. Overview of key frame animation techniques [J]. Journal of Shanxi Radio and Television University, 2009, 14(5): 55-56.

[12] 禹琳琳. 语音识别技术及应用综述[J]. 现代电子技术, 2013, 36(13): 43-45.
Yu Linlin. Speech recognition technology and application review [J]. Modern Electronic Technology, 2013, 36(13): 43-45.

[13] 侯一民, 周慧琼, 王政一. 深度学习在语音识别中的研究进展综述[J]. 计算机应用研究, 2017, 34(8): 2241-2246.
Hou Yimin, Zhou Huiqiong, Wang Zhengyi. A review of research progress in deep learning in speech recognition [J]. Application Research of Computers, 2017, 34(8): 2241-2246.

[14] 张延平, 林博文. 计算机语音集成原理、技术和应用[M]. 北京: 人民邮电出版社, 1998.
Zhang Yanping, Lin Bowen. Principles, technologies and applications of computer voice integration [M]. Beijing: People Post Press, 1998.

[15] 精英科技. 视频压缩与音频编码技术[M]. 北京: 中国电力出版社, 2001.
Elite technology. Video compression and audio coding technology [M]. Beijing: China Electric Power Press, 2001.

[16] 胡集仪. 使用 JSON 改进 WEB 数据传输[J]. 科技信息, 2008 (35): 90.
Hu Jiyi. Improve web data transmission with JSON [J]. Scientific information, 2008 (35): 90.

[17] 周丽娴, 梁昌银, 沈泽. Android 语音识别应用的研究与开发[J]. 广东通信技术, 2013 (4): 15-18.
Zhou Lixian, Liang Changyin, Shen Ze. Research and development of Android speech recognition application [J]. Guangdong Communication Technology, 2013 (4): 15-18.

[18] 陶阳. 基于 Unity 在 Android 平台上开发游戏的方法[J]. 电脑编程技巧与维护, 2012 (19): 73-77.
Tao Yang. Method of developing games based on Unity on Android platform [J]. Computer programming skills and maintenance, 2012 (19): 73-77.

[19] 付蔚, 唐鹏光, 李倩. 智能家居语音控制系统的设计[J]. 自动化仪表, 2014, 35(1): 46-50.
Fu Wei, Tang Pengguang, Li Qian. Design of smart home voice control system[J]. Automated instrument, 2014, 35(1): 46-50.

[20] 刘俐利, 凌毓涛, 王艳凤. 虚拟学习环境中构建三维动画资源与交互设计研究[J]. 中国电化教育, 2014 (2): 123-128.
Liu Lili, Ling Yutao, Wang Yanfeng. Research on Building 3D Animation Resources and Interaction Design in Virtual Learning Environment [J]. China's electrification education, 2014 (2): 123-128.

[21] 罗盛誉. Unity 5. x 游戏开发指南[M]. 北京: 人民邮电出版社, 2015.
Luo Shengyu. Unity 5. x Game development guide [M]. Beijing: People Post Press, 2015.