doi:10.3772/j.issn.1002-0470.2017.05.010

面向人机对抗赛的语音交互系统设计①

卢振利②[*,***] 田 铠[*,**] 徐惠钢[*] 张 程[*,**] 李 斌[***] 波罗瓦茨·布朗尼斯拉夫[****] 刘 军[*****]

([*]常熟理工学院电气与自动化工程学院 常熟 215500)

([**]中国矿业大学信息与电气工程学院 徐州 221116)

([***]中国科学院沈阳自动化研究所机器人学国家重点实验室 沈阳 110014)

([****]诺维萨德大学 技术科学学院 诺维萨德 21000,塞尔维亚)

([*****]浙江大学生物医学工程与仪器科学学院 杭州 310027)

摘 要 设计了中型组足球机器人的比赛中应用的人机语音交互系统。首先简要介绍了相关软件,对语音识别原理进行了解析;然后分析了语音合成技术及其实现步骤,并根据足球机器人在比赛中的实际需要,设计了一套语音指令;利用 Kinect 软件进行了实验研究;最终针对不同的发声对象测试了多组数据,实验结果表明所设计的语音交互系统对语音指令的识别行之有效,识别率较高。无论是裁判机还是队员机,都能快速准确地识别队员的语音指令并进行播报,完全满足人机对抗赛对人机语音交互的需求。

关键词 语音识别,语音合成,语音规则,Kinect

0 引 言

在中型组足球机器人的比赛中,机器人依靠助理裁判操作裁判机实现对主裁判的令行禁止,并不是像人类足球比赛一样直接听从裁判的口令。对人机对抗赛分析得出所需的信息交互手段,再结合对比人类足球比赛中,语音作为人类最基本、最自然的交流方式,是场上队员与裁判和教练进行战术交流的重要来源之一。语音技术的发展让机器人能够"听懂"人类的语言[1],因此,使用语音技术来实现人机之间的交互方式[2]是足球机器人发展的必然趋势,也是人机对抗赛中必不可少的组成部分。本研究选用微软公司的语音开发包 Microsoft Speech SDK[3]进行了语音交互的设计和开发。语音交互的基础是语音识别,微软语音识别分两种模式:文本识别模式和命令识别模式。此两种模式的主要区别在于识别过程中使用的匹配字典不同。本设计主要使用命令识别,对单个语音指令逐一测试,识别率高,速度快。在实验过程中,语音交互大大提高了足球机器人的沟通能力,能够快速识别命令,在球场上对各种情况应付自如。语音技术的发展一定会让足球机器人在未来的比赛中更加精彩。

1 语音开发工具

Microsoft Speech SDK 提供了中文的全程序引擎用于识别中文语音,是一个语音识别与合成的二次开发平台。它是基于 COM 的视窗操作系统开发工具包,含有语音应用程序接口(speech application programming interface, SAPI),微软连续语音识别(microsoft continuative speech recognition, MCSR)引擎以及串联语音合成(text to speech, TTS)引擎等等。同时它还提供了语音识别的相关组件,开放了

① 国家自然科学基金(61473283),机器人学国家重点实验室开放基金(2014-008),校新引进教师科研启动项目(XZ1306)和中国-塞尔维亚政府间科技合作委员会第 3 届例会(国科外字[2015]266 号 3-1)资助项目。

② 男,1974 年生,博士;研究方向:机器人智能控制;联系人,E-mail: zhenlilu@cslg.cn

(收稿日期:2016-11-04)

开发应用程序的接口,有可供查询的技术文档和资料,它的结构如图 1 所示。

图 1 Microsoft Speech SDK 结构

SAPI 运行库是在应用程序编程接口(API)和设备驱动接口(device driver interface, DDI)之间的,应用程序通过 API 层与 SAPI 实现语音数据处理,SAPI 则通过 DDI 调用语音引擎。它的优点[4]在于完全是基于 COM 标准开发,底层协议都以 COM 组件形式与应用程序层分离,为开发者摒除了前期复杂的语音处理过程。

语音识别技术发展到今天,特别是中小词汇量非特定人语音识别系统识别精度已经大于 98%,对特定人语音识别系统的识别精度就更高。这些技术已经能够满足通常应用的要求[5]。

2 基于 SAPI 的语音识别

语音识别技术是通过语音设备捕捉语音信号,经过一系列的处理后,语音信号转变为相应的文本或命令的技术。语音识别的处理过程从本质上来讲是一种模式识别,包含了对语音信号特征提取、基于模式匹配的识别方法和从模板库选择参考模式等三个主要的处理环节[6],它的基本原理框图如图 2 所示,最后的识别结果与语音特征、语音模型和模板的选择有着直接的关系。目前语音识别已提出了许多方法[7],模板匹配法(以动态时间归整(dynamic time warping, DTW)算法为代表)、随机模型法(以隐马尔科夫模型(hidden Markov model, HMM)为代表)和基于人口神经网络(ANN)的识别方法是比较常用的识别方法。

SAPI 拥有两种不同的语音识别引擎类型。一个是共享的语音识别引擎,可以和其他语音识别程序分享其相关程序,能提供给多个应用程序使用,所以应用广泛。另一个是独占的引擎,它只能由其创建的应用程序使用[8]。

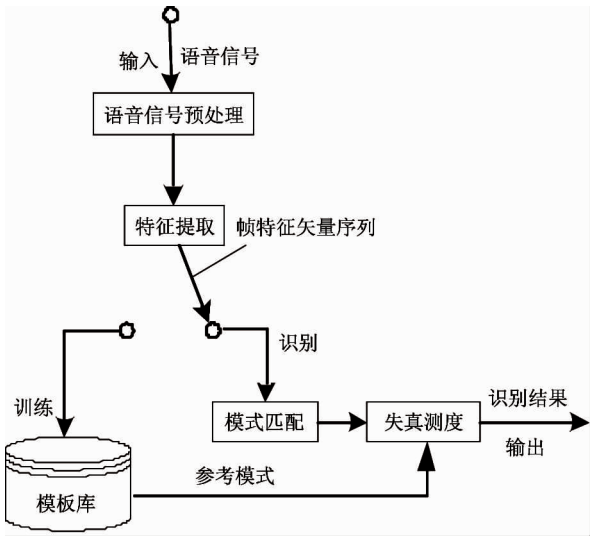


图 2 语音识别的基本原理框图

基于 SAPI 的语音识别设计方法提供的优势在于开发者不用关心采用什么样的识别方法,只需要调用对用的接口程序,便可以实现语音识别,而且开放的接口使得程序具有良好的移植性,开发效率大大提高。

本节主要对本文的语音识别方法进行设计,包括对语音识别的接口调用、语音指令规则设计和程序处理流程的设计。

2.1 语音识别接口

使用 SAPI 中提供的接口可以完成整个语音识别的处理,在 C#版的 SAPI 中,主要的识别接口为语音识别引擎(SpeechRecognitionEngine)接口、加载语法规则(LoadGrammar)接口、语音识别模式(RecognizeAsync)接口和语音识别结果(SpeechRecognized)接口。

2.2 对抗赛的语法规则设计

语音识别的语法文件有文本识别和命令识别两种模式,两种模式下分别有自己的语法规则,适合不同的场合,如表 1 所示。由于中型组足球机器人对于语音指令识别的需要是来自日常训练和比赛时裁判的命令,且都是简单的简短的语音命令,所以待识别的词汇选择命令识别模式更符合人机交互设计的需求。

表 1　两种识别模式的对比

识别模式	匹配字典	词汇量	识别速度	识别率
文本	词汇字典	大	慢	不高
命令	手动添加	小	快	高

本文将待识别的语音指令定义在一个扩展标记语言(extensible markup language,XML)文件中,与程序中的代码分离,保持语音指令的独立性。XML 可以对文档和数据进行结构化处理,实现更准确的搜索,更方便的传送,可维护性高。只需要将待识别的语音指令放在规则下,如 < Instruction item = "cyan throw-in"/ > ,程序通过加载语法文件后,就会激活规则内的语法功能。

2.3　语音识别的处理流程

语音识别的实现过程,就是先将一系列的 COM 组件初始化,加载识别引擎,设置返回消息,当语音识别事件发生后,识别引擎向程序窗口发送识别结果,然后在根据识别结果进行处理,得到所需的结果。具体的程序处理流程如图 3 所示,其步骤如下:

(1) 先对语音识别进行初始化,初始化 COM 组件。

(2) 创建语音识别对象,初始化语音识别引擎,并设定需要识别的语言。

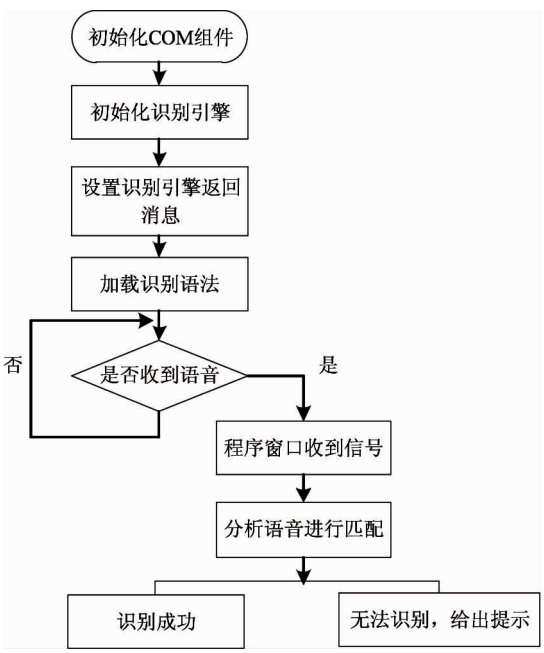


图 3　语音识别的程序处理流程

(3) 添加语音识别事件,设置语音识别引擎返回机制。

(4) 载入语法规则,等待语音事件发生,进行匹配。

(5) 程序收到语音信号后,根据语法规则和待测语音信号进行最大限度匹配,并将识别出的结果以文本形式显示。

在识别模式中,设置 RecognizeMode 的方式为 Multiple,这样就可以多次进行识别,直到关闭识别引擎或退出程序,识别过程才会结束,这样可以提高识别准确率。

3　语音合成技术

语音合成研究的目的是使机器人能够真正像人一样的说话,和人类自由地交流。将一些其他方式或存储的信息转换为语音,实现从听觉上让机器人与人进行交互。

语音合成包括 3 个主要的组成部分(见图 4)。文本分析模式对系统要处理的文本进行分词、注音,其输出是文本对应的音标序列。为了得到自然、可懂的语音输出,韵律生成模块必须对每个发音单元进行韵律调整,调整后的输出是包含韵律信息的音标序列。声学模块利用音标序列中的相应参数,选择合适的语音合成方法,生成合成语音[9]。

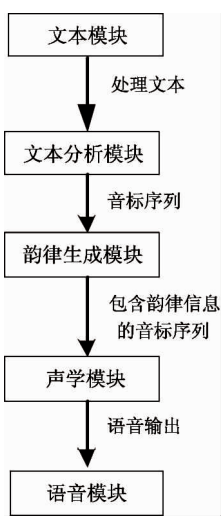


图 4　语音合成处理流程

与语音识别交互技术相比较,语音合成技术则更为成熟。现有的合成方法是基于声音合成技术[10],文本转换(TTS)是目前语音合成研究中一项优秀的技术[11],能够对文本进行实时转换,转换时间非常短,可以用秒计算。它是指计算机把文本或其他形式的信息以语音的方式输出。虽然 TTS 的自然语言程度和流畅度不是特别理想,但是合成的语音效果(清晰度和可辨识度)较好[12],速度快,能够很好地应用于人机交互中。

语音合成的具体程序步骤是:

(1)先调用 API(application programming interface)函数初始化 COM 组件。

(2)使用 SpeechSynthesizer 类创建 COM 语音合成接口实例 Ispeak。

(3)调用 Ispeak 实例中的 SpeakAsync 方法,将合成的文本字符串作为参数以异步的方式进行播放。

在播放语音之前,还可以使用 SelectVoice、Volume、Pause 等方法设置语音合成的朗读角色、音量大小、暂停等功能。

4 语音交互系统的设计

应用于本文设计的人机对抗赛的语音交互系统的设计框架如图 5 所示,首先由发声对象发出指令,输入到 Kinect,然后经 Kinect 处理生成相应的信号,输入到计算机进行语音识别,输出结果,产生的信号控制机器人运动,同时调用到 TTS 模块,经语音合成,将当前的操作或错误以语音朗读的方式播报出来[13]。

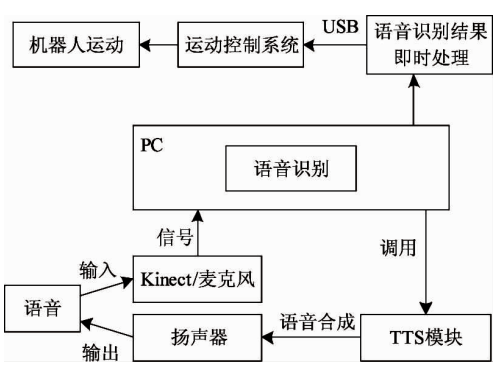


图 5 语音交互设计框架

本文设计了 2 组分别包含 20 条中、英文的文本指令,其中主要是裁判指令和操作指令。具体内容如表 2 所示,将待识别的指令存放在 XML 文件中。

表 2 语音指令列表

指令			
前进	停止	后退	左转
右转	机器人传球	机器人搜球	机器人防守
机器人抢球	机器人持球	蓝队开球	蓝队角球
蓝队边线球	蓝队点球	蓝队任意球	红队换人
红队上场	红队黄牌	红队犯规	红队进球
forward	stop	back	turn left
turn right	robot pass	robot search	robot defense
robot attack	robot holding	cyan kick-off	cyan corner- kick
cyan throw-in	cyan penalty	cyan free-kick	magenta substitution
magenta play	magenta yellow card	magenta foul	magenta goal

设计此表是为了将语音交互系统与人机对抗赛有效的连接起来,并能够针对机器人的日常训练进行人机交互,表中的指令不仅可以让机器人作为裁判使用,也可以作为队员,需根据比赛中的实际情况合理使用。

5 语音交互实验

为了得到准确的数据和真实的效果,针对本文设计的语音交互分别进行以下 4 项测试,并得出结论。

5.1 对单个语音指令测试

本次测试选用 PC 的麦克风采集声音[14],启动语音识别软件界面的识别按钮后,实验者在离 PC 的麦克风 30cm 处说出中文指令“停止”和英文指令“cyan throw-in”,得到的中文识别结果如图 6(a)所示,英文识别结果如图 6(b)所示。

除了用 PC 的麦克风作为语音输入外,还可以利用 Kinect 的麦克风接收语音指令,但 Kinect 的语音 SDK不支持中文识别,所以将Kinect的识别语音

(a) 中文识别结果

(b) 英文识别结果

图 6　语音识别的测试结果

的语言选为 en-US。Kinect 对声音的采样频率为 16kHz,采样位数为 16 位。另外,Kinect 还能对的声音来源进行定位[15]。每当实验者说出一次“forward”语音,Kinect 语音界面的识别结果如图 7 所示。

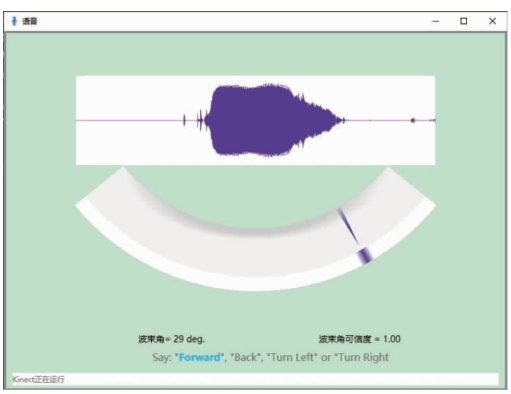


图 7　Kinect 语音测试结果

5.2　不同对象的语音识别测试

针对人机对抗赛有不同的发声对象[16],为此挑选了 3 个实验者分别对 2 种语言的语音指令进行 10 轮测试,得到的语音测试的识别率如表 3 所示。识别率 = 成功辨识次数/总试验次数 ×100%。识别结果显示出较高的识别率,识别失败的原因主要是距离麦克风过远或发音不清晰,可适当调整,提高识别率。

表 3　语音指令的识别率

指令语言	实验者 1 的识别率(%)	实验者 2 的识别率(%)	实验者 3 的识别率(%)
中文	91.5	89.5	90
英文	90.5	88	89.5

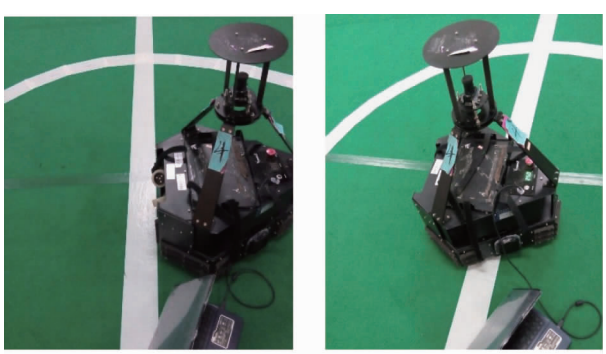
5.3　语音交互应答

利用微软的 TTS 技术,系统还设计了语音即时应答功能,主要的作用是对当前行为的播报以及操作错误时进行提醒和报警,这在人机对抗赛中可用于对队员错误指令的裁判,也可用于队员及时修正自己的错误。在系统主界面有按钮可手动选择开启,还设计了一个滑动条方便调节语音合成的音量大小[17]。实验结果显示,当实验者进行该项操作时,系统就会通过麦克风将当前的行为播放出来。

5.4　语音控制机器人实验

利用识别出的语音指令结果,可以直接与机器人的基本行为相结合。该硬件系统是通过 Kinect 麦克风阵列进行语音信号接受,在上位机进行对应指令辨识后,与下位机机器人动作控制器进行进行通讯,最后分解为万向轮速度矢量。只要语音识别成功,机器人就会执行该次命令。

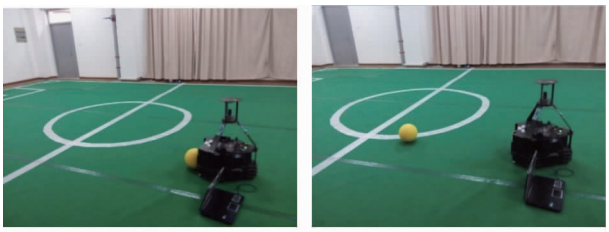
设定 $\alpha = 180$, $nLv = 300$, $\omega = 0$,当实验者以语音控制方式发出“前进”命令时,机器人由图 8(a)的位置向前运动到图 8(b)的位置。



(a) 运动前　　(b) 运动后

图 8　语音指令控制机器人前进

设置 $\alpha = 0$, $nLv = 0$, $\alpha = 0$, $v4 = 3000$, $v5 = -3000$, $nSp = 3000$, 当实验者说出语音控制命令“robot pass”后,机器人由图 9(a)的持球状态将足球弹射出去,弹射后的机器人状态如图 9(b)所示。



(a) 弹射前 (b) 弹射后

图 9 语音指令控制机器人传球

以上的实验结果显示,本文所设计的语音识别方法能够快速有效地识别指定的指令语音,应用到人机对抗赛中是完全可行的。

6 结 论

语音是人机交互中的一项重要手段,是应用到人机对抗赛中的必然趋势。本设计先对语音技术的开发工具进行了选择,进而提出本文所使用的语音识别方案,基于语音应用程序接口(SAPI)的语音识别设计,并对其中的识别接口、语法规则和程序处理流程分别进行了叙述,根据人机对抗赛的需要设计了两组中英文指令。还对语音合成技术进行介绍以及使用 SAPI 的接口设计人机交互环节的应答功能。最后,对本文设计的语音识别方法和语音交互系统进行实验,机器人在实验中表现很好,实验结果显示出本文的语音识别方法对中文和英文指令语音效果显著,并能够很好的契合到人机交互中。

参考文献

[1] 谷学静, 王志良, 郭宇承. 人机交互中的情感虚拟人技术. 北京: 机械工业出版社, 2015

[2] 李麟. 家用机器人语音识别及人机交互系统的研究: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学机电工程学院, 2007. 8-10

[3] 林茜, 欧建林, 蔡骏. 基于 Microsoft Speech SDK 的语音关键词检出系统的设计和实现. 心智与计算, 2007,1(4):433-441

[4] Zeeck A. Speech Application SDK mit ASP. NET. Springer, Xpert Press, 2005

[5] 王炳锡, 屈丹. 实用语音识别基础. 北京: 国防工业出版社, 2005

[6] Cohen I, Sebe N, Garg A, et al. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision & Image Understanding*, 2003, 91(1-2):160-187

[7] 朱淑鑫, 谢忠红. 浅谈语音识别技术的应用及发展. 长春理工大学学报:高教版, 2009, 2: 67-68

[8] 尹岩岩. 基于语音识别与合成的低速率语音编码研究:[硕士学位论文]. 上海:上海师范大学信息与机电工程学院, 2013

[9] 柳春,于洪志. 语音合成技术研究. 卫生职业教育, 2008,26(11):64-66

[10] 刘幺和, 宋庭新. 语音识别与控制应用技术. 北京:科学出版社, 2008

[11] 任鹏辉. 情感语音合成系统的研究与实现:[硕士学位论文]. 太原:太原理工大学信息与工程学院, 2013

[12] 崔斌,陈亮,胡红梅等. 基于 Kinect 的声源定位时延获取及算法性能研究. 信息技术, 2015,10:103-107

[13] 余皓,苏全. 语音控制机器人的设计与实现. 机器人技术,2007,29(5):29-31

[14] 高美娟, 杨智鑫,田景文. 移动机器人实时语音控制的实现. 电子测量技术, 2011,34(7):50-53

[15] 张梅. 隐 Markov 模型的基本原理及其在语音识别中的应用. 见:山西科技大学学报. 陕西:西安交通大学应用数学系,2003. 2-9

[16] Moscovich L G, Chen J. Learning hidden Markov models from the state distribution oracle. In: Proceedings of the International Conference on Machine Learning and Applications-Icmla, Louisville, USA, 2004. 73-80

[17] 王卫华, 顾岳. 用 Microsoft Speech SDK 实现语音识别和语音合成. 电子技术, 2000, 11:40-41

Design of a speech interaction system for man-machine confrontation

Lu Zhenli[*,***], Tian Kai[*,**], Xu Huigang[*], Zhang Cheng[*,**], Li Bin[***], Borovac Branislav[****], Liu Jun[*****]

([*] School of Electrical Engineering and Automation, Changshu Institute of Technology, Changshu 215500)

([**] School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116)

([***] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110014)

([****] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, 21000 Serbia)

([*****] Faculty of Biomedical Engineering & Instrumentation Science, Zhejiang University, Hangzhou 310027)

Abstract

A human-computer interaction system was designed by using the soccer robot game. Firstly, the related software was introduced briefly, and the speech recognition principle was analyzed; Secondly, the analysis of the speech synthesis technology and its implementation steps were performed, and according to robot soccers' actual needs in competition, a set of phonetic rules were designed; Thirdly, the experimental study on them was performed by using the Kinect software; Finally, many sets of data from different sounding objects were tested, and the testing results show that the designed speech interaction system is effective in voice instruction recognition and has the high recognition rate. The judge machine or team members in the system, can quickly and accurately identify and broadcast voice commands, showing that the system fully meets the needs of human-computer interaction in man-machine match.

Key words: speech recognition, speech synthesis, phonetic rules, Kinect