语音浏览器的概念、框架和应用

汪志鸿　张海云　陈　柯　杜利民

(中科院声学所语音交互技术研究中心　北京　100080)

摘要：分析了语音浏览器的概念和框架，以及语音输入、对话管理和响应生成整个口语对话过程的标记规范，重点说明了语音浏览器的核心标准 VoiceXML的原理和工作特性，并描述了应用语音浏览器创建口语对话系统的整体方案。以此为基础，本文给出了语音浏览器在股票交易系统中的具体应用实例。通过语音浏览器平台，可以创建灵活性很强的混合主导对话模式的口语对话系统。

关键词：语音浏览器　**VoiceXML**　口语对话系统

The Concepts Frame and Application of Voice Browser

WANG Zhihong ZHANG Haiyun, CHEN Ke, DU Limin

(Center for Speech Interaction Technology Research
Institute of Acoustics Chinese Academy of Sciences Beijing China 100080)

Abstract: The article analysis the voice browser's concepts and frame and markup standards of speech input dialog management and response output illuminate the core standard of voice browser – VoiceXML's principles and characteristics describe the whole solution of building spoken dialog system using voice browser. At last the article describe its application in the stock exchange system. We should build the mixed – interactive spoken dialog system based on voice browser.

Key words: Voice Browser; VoiceXML; Spoken Dialog System

浏览器用户主要通过阅读屏幕、操作键盘和鼠标来完成对网站的访问。下一代互联网的访问方式和网站内容的组织方式则有很大的不同。万维网各种技术标准的国际组织 W3C不仅开发了用以表达 Web页面语义的可扩展标记语言 XML，还定义了能够生成语音输出并理解语音输入的语音浏览器的技术规范。

语音浏览器支持人们以语音对话的方式与 Web页面进行交互，将扩展网站的访问人群，增加人们对网站内容的访问途径。语音浏览器的最终目标是要让人们不仅通过计算机屏幕，还能通过语音方式和互联网进行交互，将来甚至还能支持书写等其他交互方式。

语音浏览器核心的语音标记语言是建立在 XML语言基础上，实现了数据和表现的分离，计算机可以很容易地理解用户语音输入，并生成语音响应。W3C定义了一套标记语言，这些标记语言涵盖了语音识别、对话管理、语音合成、呼叫控制等语音交互应用程序的各个方面。W3C还制定了诸如语音合成标记语言、语音识别语法规范和呼叫控制 XML等规范，这些规范分别是为描述语音合成、语音识别和呼叫控制等的语法而构建的。其中，VoiceXML是对话标记语言，以它为核心，并结合语音浏览器的其他规范，就可以方便地创建系统主导或混合主导对话。

本文于 2005 – 12 – 13收到，2006 – 01 – 18收到修改稿。

1 语音接口框架

口语对话系统可以采用上述各种规范定义的标记语言来实现各个系统组件的功能,这些技术规范组成了 W 3C 语音接口框架。需要说明的是,由于不同的对话系统采用的技术不一样,组件结构也不一样,W 3C 语音接口框架并非直接定义口语对话系统中的各个系统组件,仅定义了口语对话系统中需要使用的标记语言。也就是说,语音接口框架定义的是对话系统的表示方法。

(1)VoiceXML 2 0[1]:标记语言主要用以创建语音对话,它支持口语、电话按键等输入模式和合成语音、音频文件等输出模式,也可录制数字音频,并具有对话过程控制功能。VoiceXML支持具有混合主导特征的人机对话,是开发对话管理器的有力工具。对话管理器的功能是提示用户输入,并根据用户表达的意向决定系统下一步的行为。由于对话管理器是口语对话系统中最重要的组件,VoiceXML在整个语音接口框架中也居于核心地位,后一节将对它进行更为详细的描述。

(2)SRGS[2]:语音识别的过程需要一个语言模型,识别器在语言模型给定的输入期待内寻找最可能匹配输入语音的文字序列,有效完成识别任务。采用 SRGS的语音识别器除了语音流输入外,还必须能接受用 SRGS语法形式表达的输入期待。语音识别器使用输入期待来指导语音识别过程。SRGS就针对语言模型的表达,定义了一系列的标记。这些标记可以构成特定的语法形式来表示输入期待。另外,文献[3]中定义了 n-gram 语言模型的标记语言,也可用于概念和语义的表达。

SRGS定义的标记语言称为 SGML。SGML允许使用 ABNF和 XML两种语法表达形式,这两种表达形式能够互相映射,也能自动转换。SRGS允许把 ABNF或 XML形式表达的语法嵌入到其他文档中。例如,VoiceXML2 0中的标签<grammar>支持嵌入式语法,把 ABNF形式和 XML形式的语法包含在 VoiceXML文档中。

(3)自然语言语义标记语言[4]:目前正在开发的语音识别语义解释规范定义了一种语言,可以嵌入到 SRGS语法的标签中,执行语义解释过程。语义解释处理器的输出可以使用自然语言语义标记语言表示。语义解释标签提供了一种方法,将计算语义结果的方法加入语音识别的语法中。语义解释标签处理器和 VoiceXML处理器一起使用时,能将 SRGS处理器生成的结果转换成 ECMAScript对象,然后作为 VoiceXML处理器的输入处理。

(4)SSML[5]:是语音合成标记语言,不仅用来标记生成的语音内容,还可以标记语音质量,比如音量、音速等。SSML可以和 VoiceXML配合工作,VoiceXML中<prompt>元素就是建立在以 SSML为基础的模型上,SSML中的很多元素也都可以直接用于 VoiceXML。SSML一般在口语对话系统的语言生成和语音合成组件中应用。

2 VoiceXML

VoiceXML设计目的是简化语音交互应用程序的构建过程,是一个电话语音用户接口和语音交互的技术标准。VoiceXML定义了对话结构、对话流、用户输入和系统输出等各个对话组件所需要的标记,这不仅使得 VoiceXML具备了创建语音对话的能力,而且也使得 VoiceXML在构成语音浏览器技术基础的各个标记语言中处于核心的地位。目前的 VoiceXML正式规范是 2 0版。

2 1 VoiceXML的体系结构

图 1是 VoiceXML体系结构模型图。其中 VoiceXML解释器和文档服务器之间是客户与服务器的交互关系。解释器向服务器发出请求,服务器根据请求产生需要的 VoiceXML文档作为应答发送给解释器,然后解释器负责处理收到的 VoiceXML文档。和解释器一起工作的还有解释器上下文环境组件,它的工作是监测用户的输入。

执行平台由 VoiceXML解释器环境和 VoiceXML解释器分别进行控制。语音交互应用程序中,VoiceXML

解释器环境负责检测呼叫输入,获取初始的 VoiceXML文档,并产生应答,VoiceXML解释器则负责管理应答后的对话。执行平台生成用户行为的响应事件(例如接收口语或字符输入、断开连接等)和系统事件(例如时间终止等)。

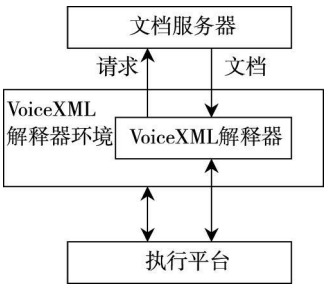

图 1　VoiceXML体系结构模型图

2.2 基本原理

VoiceXML应用程序由一组 VoiceXML文档组成,这组 VoiceXML文档中有一个根文档,在应用程序执行过程中,根文档都会一直保持,直到通话结束或者转移到另一个应用程序执行。当用户开始启动与 VoiceXML解释器交互的时候,就开始了一个完整的对话过程,在这个过程中,程序不断地装载和卸载 VoiceXML文档。在用户、VoiceXML文档或者解释器环境发出结束请求时,一个完整的对话过程才能结束。

同一个 VoiceXML应用程序中的所有文档都可以利用根文档中的信息,在执行应用程序的任何文档时,都会自动与根文档联系。一个 VoiceXML文档相当于一个对话的有限状态机,描述一系列的对话交互过程。任何时候用户总是处于某个对话状态中。每个对话通过 URI(Uniform Resource Identifier)来指定下一个对话。如果 URI没有指定一个文档,就将在当前文档进行下一个对话;如果没有具体指定下一个对话,就将指定文档的第一个对话作为下一个对话。

VoiceXML对话包括表单和菜单两类。菜单把多个可选内容提供给用户,并根据用户选择转移到另外一个对话状态。每个对话状态有一个语法相关联,语法用以描述期望的用户输入,用户输入或者是口语输入,或者是双音多频键输入。最简单的情况下,只有当前对话的语法才是激活的。在更复杂的情况下,其他语法也可以是激活的。表单定义了一个交互,这个交互收集表单中每个域的值,并提交给服务器。每个域可以指定一个提示、输入期望或评价规则。

VoiceXML还允许类似函数调用的子对话,在调用过程中保持对话的本地状态信息。子对话可被用于处理确认,也可以用于创建通用任务的可重用对话库。VoiceXML通过定义变量来保持数据。变量可以在任何一个级别上定义,按照继承型模型来确定有效范围。通过测试变量值可以决定要转移到的下一个对话状态。变量表达式也可以用在条件提示和语法中。

当提示后用户响应失败,或者当输入不能被理解时,系统将抛出事件。VoiceXML允许制定捕获事件后的处理方法。事件捕获和处理也按照一个继承模型,如果在对话级没有相应的处理方法,事件就会在更高的级别上被捕获。

VoiceXML允许你使用 ECMAScript对应用程序进行特殊控制。VoiceXML使用了类似表单填充的方法,可以在一个单一的响应中为收集多个域中的值定义一个复杂的语法,任何未填充的域都可以由每个对话内定义的特殊子对话处理。

2.3 混合主导

VoiceXML不仅能够创建系统指导的对话,还能创建混合主导对话。混合主导对话是指用户和系统都可以主导对话进程,而系统指导的对话流程是由系统预先设定好的。与系统指导的对话相比,混合主导的对话界面更友好。由于混合主导的对话每个对话回合都可以表达更多的信息,设计得好的混合主导对话能够用较少的回合完成同样的对话任务。VoiceXML混合主导的特点是表单驱动的。

4 股票交易系统的语音浏览器解决方案

语音浏览器作为新一代浏览器,尽量将复杂的人机对话系统构建过程实现简单化和模块化。本节描述了一个基于语音浏览器开发平台的股票交易系统。其基本框图如图 2所示。

作为一种电话语音系统,其最常见的应用领域是信息的查询与获取、操作指令的理解与执行等服务。

语音股票交易系统基于语音浏览器的一般原理，系统基本构成包括语音识别组件、语言理解组件、对话管理组件、语言生成组件和语音合成组件。

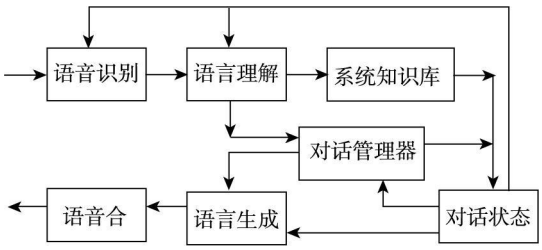

图 2 混合主导口语对话系统

语音识别组件：通过使用 SGML标记的语言模型对用户的口语输入进行语音识别。

语言理解组件：使用预先指定的语法从文本字符串中抽取语义信息。文本字符串可由语音识别器产生，也可以是用户通过键盘输入。理解过程也使用了 SGML标记信息，并对语言理解的结果进行标记。通过自然语言标记，可以得到用户查询或者买卖股票等意图信息和股票名称等语义概念。

对话管理组件：主要任务是提示用户输入，并根据用户表达的意向决定系统的响应行为，使用 VoiceXML标记对话脚本，可以设计混合主导对话。根据用户的查询指令或者买卖指令完成相应的操作。

语言生成组件：产生响应用户的文本。文本可以包含使用语音合成标记语言标签。这些标签说明文本如何发音。

语音合成组件：将语言生成组件产生的响应文本转化为声音。基于语音浏览器平台构建的系统是由口语对话系统中领域无关模块构成的，这部分对于各具体任务领域来讲是通用的，具备可移植性。因此只要基于语音浏览器平台定义好领域相关模块，在该平台上就可开发出针对具体任务领域的各种口语对话系统。这样可以最大限度地提高口语对话系统的开发效率和可移植性。

5 结束语

语音浏览器定义了理解语音输入、响应语音生成等整个口语对话过程的标记规范，其最终目标使人们能够能通过语音和万维网进行直接交互。本文分析了语音浏览器概念和框架，并仔细描述了其核心 VoiceXML对话标记语言部分。以此为基础，给出了语音浏览器在股票交易系统中的具体应用实例，通过它可以创建股票交易中的混合主导对话模式。语音浏览器有着巨大的应用潜力，随着研究的发展，它将给越来越多的领域带来深刻的变化。

参 考 文 献

1 Scott McGlashan, Daniel C. Burnett, Jerry Carter, Peter Danielsen, Jim Ferrans, Andrew Hunt, Bruce Lucas, Brad Porter, Ken Rehor, Steph Tryphonas. Voice Extensible Markup Language (VoiceXML) Version 2.0. http://www.w3.org/.

2 Andrew Hunt, Scott McGlashan. Speech Recognition Grammar Specification Version 1.0. http://www.w3.org/.

3 Michael K. Brown, Andreas Kellner, Dave Raggett. Stochastic Language Models (N-Gram) Specification. http://www.w3.org/.

4 Deborah A. Dahl. Natural Language Semantics Markup Language for the Speech Interface Framework. http://www.w3.org/.

5 Daniel C. Burnett, Mark R. Walker, Andrew Hunt. Speech Synthesis Markup Language (SSML) Version 1.0. http://www.w3.org/.

作者简介

汪志鸿，男，(1975-)，博士生，主要研究方向：自然语言理解、口语对话系统。

张海云，男，(1972-)，博士生，主要研究方向：语音信号处理与语音识别。

陈 柯，男，(1977-)，博士生，主要研究方向：双模态语音交互技术。

杜利民，男，(1957-)，研究员，博士生导师，主要研究方向：语音信息交互处理技术。