제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

음성 인식 사용자 인터페이스를 통한 가전기기 제어 기법*

송욱, 장현수, 엄영익
성균관대학교 컴퓨터공학과
e-mail:woogilord@skku.edu, {jhs4071, yieom}@ece.skku.ac.kr

Home Appliance Control through Speech Recognition User Interface

Wook Song, Hyun-Su Jang, Young-Ik Eom
School of Information and Communication Engineering,
Sungkyunkwan University

요 약

유비쿼터스 컴퓨팅 환경이 확대됨에 따라, 기존의 키보드와 마우스만을 사용자 인터페이스로 주로 사용했던 방법에서 벗어나 좀 더 사용자 중심의 멀티모달 유저 인터페이스 적용이 요구되고 있다. 이에 XHTML+Voice는 음성 및 시각을 모두 제공할 수 있는 새로운 서비스 패러다임으로서 기존의 음성정보만을 제공하거나 시각적인 정보만을 제공하는 시스템과는 달리 XHTML내에 VoiceXML을 삽입함으로써 두 언어의 장점을 모두 활용할 수 있다. 본 논문에서는 VoiceXML의 이러한 장점을 살려 스마트 홈을 구성하는 여러 가전기기들의 인터페이스를 미리 템플릿으로 만들어 두어 모바일 디바이스를 통해 이것들을 제어하는 시나리오를 제안하고 구현하는 방법에 대해 실험하였다.

1. 서론

다양한 입출력 장치를 통한 사용자와 시스템과의 상호작용은 현재 사용하고 있는 GUI 등의 특정 입출력 장치만을 이용하는 방식보다 인간과 인간이 대화하는 방식에 가까운 환경을 제공해준다. [1] 멀티모달 유저 인터페이스 적응(Multimodal UI Adaptation)은 사용자의 속성 및 습관, 서비스를 제공 받을 기기의 구동 환경, 사용자에 대한 정보 등을 바탕으로 가장 적절한 모달리티들을 갖는 인터페이스를 제공한다. 현재 유비쿼터스 컴퓨팅 환경을 지원하기 위한 여러 미들웨어들의 개발이 활발하게 진행되어, 이러한 미들웨어를 탑재한 가전기기들이 머지않아 시장에서 주류가 될 것이다. 그렇게 되면 모바일 장치를 이용하여 각 장치의 특성에 맞는 유저 인터페이스를 사용자가 원하는 모달리티로 제공받아 제어하는 홈 리빙 서비스 역시 대중화 될 것이다. 본 논문에서는 음성인식이 가능한 모바일 디바이스를 통해 홈 게이트웨이 서버에 인증을 한 후, 사용 가능한 가전기기의 UI를 XHTML+Voice로 제공받아 그것을 음성과 간단한 제스처로 제어하는 서비스 시나리오와 응용 어플리케이션의 구조를 소개한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 본 논문에서 제시하고자 하는 시스템의 주가 되는 VoiceXML과 XHTML, X+V에 대해 소개하고, 3장에서는 음성을 통한 가전 기기를 제어에 관련한 제안 사항을 소개할 것이며 4장에서는 제안 사항에 대한 실험과 구현에 대한 소개를 할 것이다.

2. 관련 연구

현재 사용되고 있는 HTML은 근본적인 한계로 인해 응용 및 활용에 어려움을 겪고 있다. HTML의 단점을 보완하여 XML이 등장하였다. VoiceXML과 XHTML은 XML의 유연한 확장성으로 인해 생긴 파생 언어의 한 종류라 볼 수 있다.

2.1. VoiceXML

XML은 W3C(World Wide Web Consortium)에서 HTML의 한계를 극복하기 위한 대안으로 만든 마

* 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (IITA-2005-(C1090-0501-0019)).

크업 언어로써 플랫폼에 독립적이며 무한한 확장성을 통해 다양하게 접목될 수 있다는 장점을 가지고 있다. 음성이라는 모달리티에 한정하여 기존의 음성 서비스는 인터넷의 광범위한 서비스를 활용하는 데에 있어서 시나리오를 작성하는 표준 언어의 부재로 같은 시나리오도 서비스 플랫폼에 따라 매번 새로 제작해야 한다는 단점을 갖고 있었다. 그러나 Voice XML은 음성 서비스 시나리오 작성에 대한 표준으로 제시되고 있으며 매우 보편화 된 음성 전화를 사용하여 웹 연결을 가능하게 하는 메커니즘을 제공한다.

2.2. XHTML

XHTML(eXtensible HyperText Markup Language)은 HTML와 완전히 개별적인 언어이다. XML이 일반 웹브라우저에서 읽을 수 없는 점을 감안해서 일반 웹브라우저를 통해서도 접근할 수 있도록 HTML 태그들을 공유할 수 있는 웹페이지 마크업 언어로서 W3C의 차세대 표준안이다.

2.3. X+V (XHTML+Voice)

XHTML+Voice는 XHTML의 모듈화로서 기능이 확장된 XHTML의 한 종류이다. XHTML에서 사용되는 입력 및 텍스트 엘리먼트와 VoiceXML의 다이얼로그 사이의 모달리티를 공유하기 때문에 두 마크업 언어가 제공하는 시각, 음성 모달리티를 모두 사용하여 사용자와 상호 작용이 가능하다.

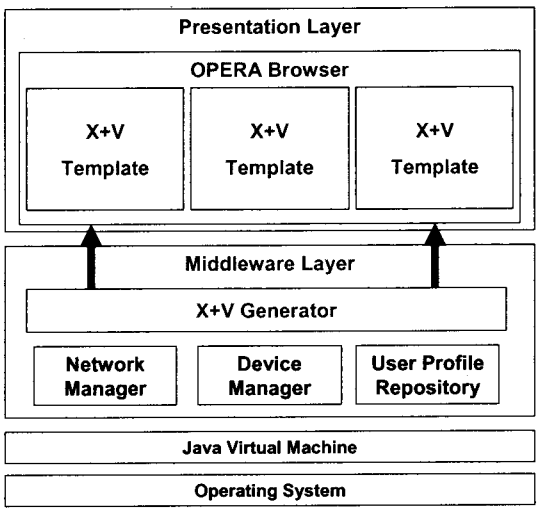
3. 제안

3.1. 제안 시스템 아키텍처

본 논문에서는 (그림 1)과 같은 형태의 시스템 아키텍처를 제시한다. 그림에서 알 수 있듯이 시스템은 크게 Operating System, Java Virtual Machine, Middleware, Presentation Layer의 4개의 부분으로 구성되어 있다. Middleware Layer에는 Network Manager, Device Manager, User Profile Manager, X+V(XHTML+Voice) Generator의 컴포넌트를 갖고 있다. Network Manager는 스마트 홈을 구성하는 여러 가전기기와의 통신을 담당하고 이 과정에서 오고가는 메시지를 분석, 처리한다. Device Manager는 여러 가전기기의 장치 정보를 관리하는 컴포넌트로써 각 기기가 갖는 공통적인 특성을 자체로 관리하고 있다. 예를 들어 TV라는 가전기기가 있을 때, TV는 기본적으로 채널, 볼륨 조정과 같은 인터페이스를 갖고 있는데 이러한 정보를 미리 XML 문서로 만들어서 관리한다. User Profile Repository는 적응적 유저 인터페이스 구성을 위한 사용자의 기호 등을 저장해 두는 곳으로서 역시 X+V Generator에게 전달되어 XHTML+Voice 코드를 생성하는데 사용된다.

X+V Generator는 Presentation Layer에서 브라우저를 통해 사용자에게 보여 지는 유저 인터페이스를 생성하는 컴포넌트이다. 유저 인터페이스는 XHTML +Voice 문서로 생성되며, 생성 과정에서 적응적 유저 인터페이스 생성을 위해 Device Manager와 User Profile Repository의 정보를 활용하게 된다.

Presentation Layer에서는 사용자에게 실제로 사용자에게 인터페이스를 제공하는 역할을 담당한다. X+V Generator에서 생성된 XHTML+Voice 문서는 OPERA 브라우저[5]를 통해 사용자에게 보여 진다.



(그림 1) 음성 인식 사용자 인터페이스 제공 시스템

3.2. 유저 인터페이스 예제 스크립트

아래 (그림 2)는 텔레비전을 조작하는 음성 인식 사용자 인터페이스의 XHTML+Voice 스크립트에서 볼륨을 조정하는 부분만을 발췌한 것이다. 볼륨은 간단한 텍스트 입력창을 통해서 입력받는 형식으로 터치펜 등을 사용하여 텍스트 입력창에 커서가 가게 되면 VoicxXML로 작성된 이벤트 핸들러가 처리하게 된다. 위의 스크립트에서 page.output_box의 아이디를 갖는 입력폼에 커서가 가게 되면 vxml_volume_prompt 라는 아이디를 갖는 VoiceXML 폼이 이 이벤트를 받아 처리하게 되어 사용자에게 원하는 볼륨을 음성으로 입력해 줄 것을 요구하는 음성 안내와 현재의 볼륨을 사용자에게 알려준다. 0에서 30까지의 볼륨을 사용자의 음성을 통해 입력 받게끔 JSGF [8] 문법을 정의하여 음성을 장치가 인식할 수 있도록 하였다. 여기서 초기 볼륨값과 0에서 30사이의 볼륨 범위는 타깃 장치에 의존적일 수 있으므로 이러한 정보는 위에서 설명하였던 시나리오에서처럼 User Profile Repository와 Device Manager를 통해

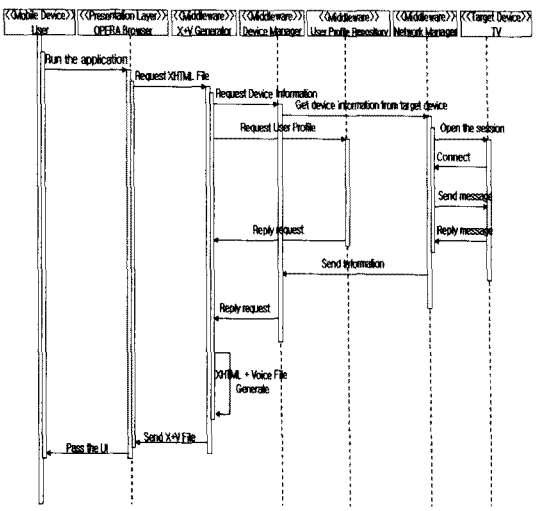
획득한 정보를 사용하여 적응적 유저 인터페이스를 구성한다.

```
<vxml:form id="vxml_volume_prompt">
<vxml:field name="vxml_volume_field">
<vxml:grammar>
<![CDATA[
#JSGF V1.0;
grammar volume_selection;
public <volume_selection> =
0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |12 | 13 | 14| 15 | 16 |
 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26| 27 | 28 | 29 | 30; ]]>
</vxml:grammar>
<vxml:prompt>
Please tell me the level of sound volume that you want to adjust between 0 and 30.
Current volume level is <vxml:value expr="volume_var"/>
</vxml:prompt>
<vxml:filled>
<vxml:assign name="volume_var" expr="vxml_volume_field"/>
<vxml:assign
 name="document.getElementById('page.output_box')..value"  expr="volume_var"/>
</vxml:filled>
</vxml:field>
<vxml:block>
Now the volume level changes to <vxml:value expr="volume_var"/>
</vxml:block>
</vxml:form>
.
.
.
<TABLE BORDER=5 >
<TR><TH> TV Control Box </TH>
<TD>
Volume : <input type="text" id="page.output_box" ev:event="focus"
 ev:handler="#vxml_volume_prompt" value="" size="18"/>
```

(그림 2) TV 제어를 위한 음성 인식 사용자 인터페이스 스크립트

3.3. 시나리오

다음 (그림 3)은 TV를 제어하기 위한 시나리오의 시퀀스 다이어그램이다.



(그림 3) TV 제어 시나리오

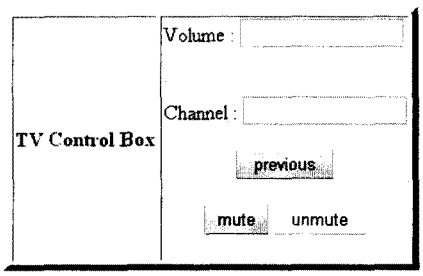
위의 시나리오에서는 모바일 장치를 사용하여 보편적인 가전기기인 텔레비전을 조작하기 위한 유저 인터페이스를 확보하는 과정을 간단히 소개한다. 본 시나리오에서는 사용자가 로그인하여 사용자 인증을 마친 상태라 가정한다.

1) 사용자는 자신의 모바일 장치에서 타깃 장치를 제어하기 위해 응용 프로그램을 구동한다.
2) 응용 프로그램은 X+V Generator에게 유저 인터페이스가 구성되어 있는 XHTML+Voice 파일을 요청한다.
3) X+V Generator는 Device Manager에게 조작하기를 원하는 타깃 장치의 정보와 사용자의 정보를 요청한다.
4) Device Manager는 응용 프로그램이 구동되고 있는 장치의 장치 정보를 X+V Generator에게 전달한 후 Network Manager를 통해 타깃에 접근한다.
5) Network Manager는 약속된 메시지를 전달하여 원하는 정보가 담긴 답신을 받아 장치 정보를 Device Manager에게 넘겨준다.
6) Device Manager는 타깃 장치만이 갖고 있는 특별한 인터페이스 정보를 추출하여 약속된 형식으로 X+V Generator에게 넘겨준다.
7) X+V Generator는 템플릿 파일에 타깃 장치에서 추출한 특별 인터페이스 정보와 사용자 속성 정보를 옵션의 형태로 업데이트한다.
8) X+V Generator는 X+V 파일을 생성하고, 브라우저를 통해서 사용자에게 넘겨준다.

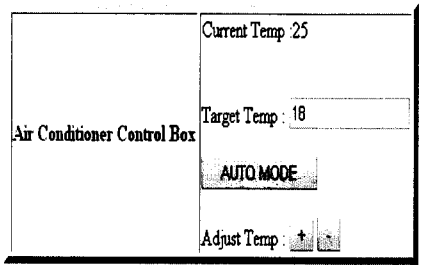
4. 구현 및 평가

본 시스템은 인텔의 PENTIUM4 3.0Ghz CPU와 512MB의 DDR(PCI2700)의 사양을 갖는 일반적인 데스크탑 환경에서 운영체제는 Windows XP, XHTML+Voice를 보여주기 위한 브라우저는 OPERA를 사용하여 제작, 실험되었다.

4.1. 가전기기 제어 유저 인터페이스



(a) 텔레비전을 조작하기 위한 UI



(b) 에어컨을 조작하기 위한 UI

(그림 4) 가전기기 제어 유저 인터페이스

위의 (그림4)는 본 시스템에서 사용자에게 보여주게 되는 최종 유저 인터페이스의 모습이다. (a)의 인터페이스는 텔레비전은 조작하기 위해서 생성한 유저 인터페이스고 (b)의 인터페이스는 에어컨을 조작하기 위한 인터페이스이다. 텔레비전을 조작하기 위해 필요한 최소한의 기능으로써 볼륨 조정, 채널 조정, 이전 채널, 소리줄임, 소리줄임 해제를 조작할 수 있도록 버튼을 디자인 하였다. 모든 버튼 및 입력 폼은 커서를 얻을 때 마다 음성으로 어떤 기능을 하는지에 대한 설명을 하게 되고, 사용자는 그 음성 메시지에 맞추어 음성으로 제어가 가능하다. 에어컨 역시 텔레비전과 마찬가지로 희망 온도 설정, 자동 동작 모드, 희망 온도 증가 감소의 기능만을 갖는 인터페이스를 사용자에게 보여준다.

4.2. 평가

본 시스템의 궁극적인 목적인 모바일 장치에서의 테스트는 아직까지 XHTML+Voice를 사용자에게 보여줄 수 있는 모바일 장치용 브라우저가 존재하지 않으므로 시도하지 못하였으나, XML을 기반으로 한 XHTML이므로 구동 시스템에 상관없이 동작하리라 예상된다. 추후에 XHTML+Voice를 지원하는 브라우저를 이용해서 실험할 계획을 갖고 있다.

5. 결론

유비쿼터스 컴퓨팅 환경이 확대됨에 따라 사용자 중심의 멀티모달 유저 인터페이스의 요구가 증가하고 있다. 이러한 추세에 따라 본 논문에서는 XHTML+Voice을 이용하여 음성 및 시각 모달리티를 모두 제공하는 가전기기 제어 유저 인터페이스를 제안하였다. 미리 제작한 템플릿에 옵션 형식으로 장치 정보와 사용자 정보를 설정하고 타깃 장치의 인터페이스를 그대로 가져오지는 못하기 때문에 아직까지 사용자 인터페이스에 많은 기능을 제공하지는 못하였다. 향후 타깃 장치의 장치 정보와 기능을 XML로 상세히 구술하여 만들어진 XML을 템플릿에 효과적으로 적용하는 방법에 대한 연구를 진행할 예정이다.

참고문헌

[1] Carlos Duarte and Luís Carriço, "A conceptual Framework for Developing Adaptive Multimodal Applications", Proc. of the 11th international conference on Intelligent user interfaces, SESSION: Multimedia and Multimodality, Sydney, Australia, pp. 132-139, 2006.

[2] Oliver Lemon, Anne Bracy, and Alexander Gruenstein, "The WITAS Multi-Modal Dialogue System I", Proc. of European Conference on Speech Communication and Technology, pp. 1559-1562, Sep. 2001.

[3] X+V: Multimodal App, http://www-306.ibm.com/software/pervasive/multimodal/

[4] VoiceXML, http://www.voicexml.org/

[5] Multimodal Browser, http://www.opera.com/products/devices/multimodal/index.dml

[6] XHTML+Voice Profile 1.2, http://www.voicexml.org/specs/multimodal/x+v/12/spec.html

[7] XHTML™ 2.0, W3C Working Draft 26 July 2006 , http://www.w3.org/TR/xhtml2/

[8] Java™ Speech Grammar Format Specification, http://java.sun.com/products/java-media/speech/forDevelopers/JSGF.pdf