유비쿼터스 환경의 정보서비스를 위한 음성기술 표준화 동향

성신여자대학교 홍기형
서울대학교 정민화

1. 서 론

언제, 어디서, 어떤 장치를 사용하더라도 정보의 검색이나 접근이 가능해지는 유비쿼터스 환경이 가시화되고 있어, 음성을 사용한 사용자 인터페이스의 중요성이 증대하고 있다. 음성정보기술은 인간의 가장 자연스러운 상호작용 수단인 음성을 이용하여, 시스템에 명령을 내리고, 시스템의 명령 수행 결과를 음성으로 전달하는 음성을 이용한 정보 시스템 인터페이스를 구현하기 위한 기술을 의미한다. 이동 중과 같이 모니터 등 시각적 인터페이스의 사용이 용이하지 않은 상황에서 정보 접근의 요구가 증대함에 따라 음성은 정보시스템의 중요한 인터페이스로 자리매김하고 있다. 음성정보시스템은 음성을 이용한 사용자인터페이스가 가능한 정보시스템을 뜻하며, 그림 1과 같이 음성처리엔진, 사용자 접속망, 사용자 프로파일 및 음성응용시스템으로 구성된다.

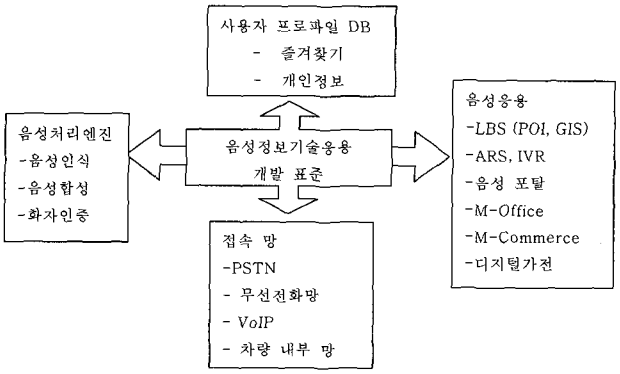

그림 1 음성정보시스템 구성 및 표준

음성정보기술의 응용분야는 기존의ARS(Automatic Response System), IVR(Interactive Voice Response) 시스템 에서, 자동차용 텔레매틱스, 디지털 가전 및 홈 오토메이션, 로봇 등으로 확대되고 있으며, 이에 따라 다양한 응용에 음성정보기술을 적용하기 위한 편리한 응용 개발 환경과 음성 응용 프로그램의 플랫폼 독립성 및 호환성의 확보가 절실히 요구되고 있다. 이에 따라, 음성정보기술 표준은 그림 1에서 보인 바와 같이 음성정보시스템을 구성하는 4 가지 요소의 상호 독립성을 보장하여야 한다. 여기서, 음성정보기술 응용, 또는 음성응용이란 디스플레이 장치를 통하여 눈으로 보고, 키보드로 입력하는 기존의 정보시스템의 그래픽 사용자 인터페이스(*GUI: Graphic User Interface*)를 귀로 듣고, 말하는 음성 사용자 인터페이스(VUI: Voice User Interface)로 대체한 것을 말한다.

음성정보기술 표준은 각기 다른 기관이 개발한 음성인식/합성 시스템 등 원천 기술 기반 시스템, 다양한 접속 망, 서로 다른 정보시스템(Back-end Data Server 및 Back-office 등) 등을 개발하고자 하는 음성 응용과 분리하여 생각할 수 있도록 한다. 음성응용 개발을 위해 제정된 표준에 따라 개발된 음성 응용은 각기 다른 기관에서 개발한 음성 인식/합성 시스템에서 별도의 수정 작업 없이 그대로 사용할 수 있도록 한다. 다시 말하면, 표준에 따라 개발된 음성 응용 S/W는 그 표준을 지원하는 음성 인식/합성 시스템이라면, 어느 기관에서 개발한 것이든 상관없이 수정 없이 그대로 적용할 수 있다는 것을 뜻한다. 뿐만 아니라 음성응용에서 개인화 서비스를 위한 각종 프로파일 정보에 대한 접근 방법, 각종 망을 통한 음성시스템 접근 방법도 표준화하여, 응용 개발의 편리성을 보장하고 응용의 확대를 꾀하고 있다.

음성정보기술의 표준은 초기에는 음성인식/합성 등 원천기술을 위한 응용프로그래밍 인터페이스(API)의 표준화부터 시작하였다. 이러한 음성 엔진 API로는 Java 기반으로 JSAPI(Java Speech API)[1], JTAPI(Java Telephony API) [2] 등이 있으며, 마이크로 소프트 운영체제에서는 MS SAPI(Speech API) [3] 등이 개발 되었다. 이들 API 수준의 표준은 음성 인식/합성기와 전화망 제어 기능을 표준화된 API로 분리함으로써, 음성정보시스템 개발자는 이들 API 만을 숙지하면, 이를 이용하여 음성 대화 시나리

오를 개발한다. 그러나, 이러한 경우에도, 음성인식/합성 엔진의 특성이나 전화망의 특성 등을 이해하지 못하면, 이들 API 자체의 이해가 불가능하고, Java나 C, C++프로그램이 가능한 개발자이어야 음성 대화 시나리오를 개발할 수 있다.

90년대 초에 등장한 Web은 이전의 정보 시스템에서의 개발 환경, 사용자 인터페이스 등에 많은 영향을 미쳤다. 특히, 웹 브라우저는 예약, 뱅킹 등을 비롯한 모든 정보시스템의 사용자 인터페이스로 자리 잡고 있으며, HTML, XML로 대표되는 사용자 인터페이스 개발 언어는 개방형 표준으로 이전의 정보시스템 인터페이스 개발 환경에 비하여 많은 장점을 가지고 있다.

- 개방형 표준 GUI 기술 언어(HTML) : HTML이란 표준 GUI 기술 언어가 있다. 예약이나 뱅킹과 같은 정보 시스템의 사용자를 위한 GUI 인터페이스 기술 언어인 HTML은 C, C++, Java 등의 프로그래밍 언어에 비하여 배우기가 매우 쉽다.
- GUI 해석기인 웹 브라우저 : HTML로 기술된 GUI는 웹 브라우저라고 하는 HTML 해석기가 있어 HTML로 표현된 GUI를 사용자의 모니터에 실현해 준다. GUI 개발자는 마크업 언어인 HTML로 원하는 GUI 인터페이스를 설계하고 작성한다.
- 실시간 자동 HTML(GUI) 작성 : CGI(Common Gateway Interface) 나 ASP(Active Server Page) 등을 이용한 웹서버 쪽의 프로그램으로 정보시스템의 상황에 따라 다름 형식의 HTML 문서가 작성될 수 있다. 이러한 CGI, ASP 프로그램을 이용하면, 특정 웹 사이트에 접속하는 사용자 마다 해당 사용자에게 맞춤형의 HTML 문서를 통하여 맞춤형 GUI를 제공할 수 있다.
- 다른 정보시스템과의 연결 용이 : 하이퍼텍스트 개념을 기본으로 하는 Web 환경에서는 다른 기관, 다른 플랫폼에서 제공하는 정보시스템을 접근하기 위하여, 링크(link)를 둘 수 있어, 사용자는 특정한 정보 시스템 사이트로 접근하여, 링크되어 있는 다른 정보 시스템에 쉽게 접근 할 수 있다.

이러한 장점으로 인하여, 현재 웹은 거의 모든 정보시스템의 인터페이스 개발 및 운영 환경으로 자리 잡고 있다. 그래픽 모니터의 특성이나, 정보 시스템과의 인터페이스에 대한 이해가 없어도 간단한 HTML 마크업 만을 알면, 초등학생이라도 작성할 수 있으므로, 이에 따라 HTML의 확산이 매우 빠르게 진행되었다.

90년대 후반에 음성정보기술 개발에 있어서도 이러한 웹의 개발환경 및 개방형 표준을 접목하기 위한 연구가 시작되었고, 그 결과, W3C에서는 음성 브라우저(Voice Browser) 표준화 활동 [4]이 계속되고 있다.

최근에는 사용자와 시스템 사이의 상호작용(인터페이스) 정확성을 보다 향상시키고, 보다 인간이 자연스러운 인터페이스를 추구하기 위하여, 음성을 기반으로 한 멀티모달 인터페이스를 위한 국제 표준화 작업이 진행되고 있다. 멀티모달 인터페이스는 음성, 그래픽(GUI), 키보드, 키패드, 펜(필기 인식) 등 다양한 입출력 방법을 동시에 사용하는 인터페이스 기술을 말한다. 이러한 멀티모달 인터페이스와 관련한 표준은 정보시스템 플랫폼에서 전 세계 시장의 상당 부분을 차지하고 있는 마이크로소프트사가 주축이 되어 제안하고 있는 SALT(Speech Application Language Tags) [6]와 IBM이 주축이된 XHTML+Voice [7]가 있고, W3C에서 진행하고 있는 멀티모달 상호작용 표준화가 있다.

모바일 환경에서 사용자가 사용하는 휴대전화 또는 PDA 등의 소형 단말기의 성능이 현재의 음성 인식/합성 기술에서 일정한 질(quality)를 보장하기 위한 하드웨어 요구사항을 만족하기가 어려운 상황에서 단말에서는 음성 신호에 대한 디코딩/엔코딩만을 담당하고 음성 인식 및 합성은 보다 하드웨어 성능이 뛰어난 서버 기반으로 수행하는 분산음성처리를 위한 표준화 [9,10]가 진행되고 있다. 특히 이러한 분산 음성 처리 표준은 유비쿼터스 환경에서 음성 인터페이스의 활성화를 위하여 매우 중요하며, 유무선 인터넷 기반의 MRCP(Media Resource Control Protocol) [10]은 사용 기기에 독립적인(device independency) 음성 응용 개발에 매우 중요한 표준이다.

본 고에서는 **음성정보기술과 관련한 국제 표준화 동향**에 대하여 알아보고자 한다. 2장에서는 웹 개발 환경을 음성응용 개발에 적용하기 위한 W3C(World Wide Web Consortium)의 음성 브라우저(Voice Browser)관련 표준을 알아보았다. 3장에서는 음성를 포함하는 멀티모달 인터페이스 표준화 동향을 기술하였다. 4장에서는 단말에서 서버기반의 음성 엔진을 이용한 음성 정보 시스템 개발을 용이하게 하는 MRCP에 대하여 알아보았다. 마지막 결론으로 향후 음성 응용 개발 표준의 방향에 대하여 기술하고자 한다.

2. Web 기반 음성 기술 표준화

그림 2는 W3C 음성 브라우저 [4]의 구조 및 각 요소 별 표준화 명세를 보이고 있다. 웹 기반 음성대화 관리, 음성 합성, 음성 인식, 전화망 제어(call control), 그리고 기타 쌍방향 음성 응답 응용을 포함하는

마크업(markup) 언어 집합을 규정하고 있다. SSML (Speech Synthesis Markup Language), SRGS (Speech Recognition Grammar Specification), CCXML(Call Control XML)와 같은 명세서들은 각각 음성의 합성과 인식 문법, 그리고 전화망 제어를 기술하기 위한 언어 명세이다. VoiceXML은 합성 음성, 오디오, 음성 & DTMF key(touch tone) 입력의 인식, 전화 통신을 기반으로 하는 대화(dialog)를 기술하기 위한 음성 대화 기술 마크업(dialog markup) 언어이다. 이러한 XML 기반의 음성 대화 및 인식/합성 마크업은 웹 기반 컨텐츠의 음성 서비스를 가능하게 할 뿐 아니라, 웹 기반 개발의 장점을 음성정보 응용 개발에서 가능하도록 한다.

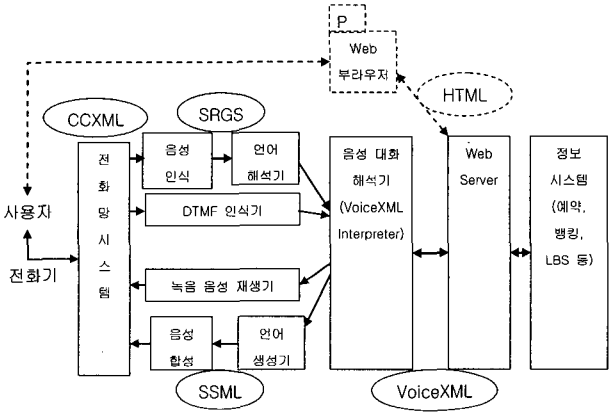

그림 2 W3C 음성 브라우저

표 1은 W3C 음성 브라우저 관련 표준의 현황이다. 여기서, WD는 working draft로 표준이 진행 중인 것을 의미하며, 이중에서 last call이란 표준(recommendation)으로 제정하기 바로 직전 단계의 표준 문서임을 의미한다. CR은 Candidate Recommendation으로 표준 후보 안으로 확정된 것을 의미한다.

W3C 음성 브라우저 관련 표준 개발은 시급성의 여부에 따라 표준화 항목을 다음과 같이 나누어 진행되고 있다.

- 고 순위 표준화 항목 : dialog(VoiceXML2.0), speech recognition grammar(SRGS), speech synthesis(SSML), semantic interpretation, call control(CCXML)
- 저 순위 표준화 항목 : pronunciation lexicon, stochastic grammars(N-Grams), voice browser inter-operation

저 순위 표준화 항목은 고 순위 항목의 표준이 완성된 다음에 다시 시작될 것 같다. 또한, 2003년 초반부터, 음성 대화 기술 마크업 언어의 차기 버전을 위한 요구 사항 수집을 시작 했으며, 멀티모달 인터페이스 응용에서 음성을 지원하기 위한 활동을 사작하고 있다.

표 1 W3C 음성 브라우저 표준화 현황(2005년 11월 현재)

분 류	표 준	현재 상태
Voice XML	VoiceXML 2.1	CR, 13 June 2005
	VoiceXML 2.0	Recommendation, 16 March. 2004
Speech Synthesis	Speech synthesis requirements	WD, 23 December 1999
	Speech synthesis specification	Recommendation , 8 September 2004
Speech Recognition	Speech grammar requirements	WD, 23 December 1999
	Speech Grammar specification	Recommendation, 16 March. 2004
Semantic Interpre-tation	Semantic Interpretation for Speech Recognition	CR, expected December 2005
Pronunciation Lexicon	Pronunciation lexicon requirements	WD, 14 February 2005
Call Control	Call Control XML(CCXML)	Last Call Working Draft, 29 June 2005

각 표준화 항목의 개발 현황을 요약하면, 다음과 같다.

2.1 VoiceXML

- Requirements(요구분석)-1999년 12월 23일
- W3C VoiceXML 2.0 Recommendation-2004년 3월 16일
- VocieXML 2.1 Candidate Recommendation -2005년 6월 13일

VoiceXML 2.0은 합성음, 디지털 오디오, 음성인식, 음성 녹음, DTME 인식을 바탕으로 전화망 환경에서 상호 주도형 대화(mixed-initiative conversation)을 기술할 수 있도록 설계 되었다. VocieXML 2.1은 이동형 음성 기기를 위한 VocieXML 2.0의 확장이며, 현재 다른 웹기반 언어와의 통합을 위한 VocieXML 3.0이 준비되고 있다.

2.2 Speech Recognition Grammar Specification (SRGS)

- Requirements-1999년 12월 23일
- Candidate Recommendation-2002년 6월 26일
- Recommendation - 2004년 3월 16일

SRGS는 음성과 DTMF(touch tone) 입력 문법을 둘 다를 기술할 수 있는 표준이다. 일반적으로 유/무선

전화의 키패드에서 입력되는 DTMF는 시끄러운 상태나 문맥상 말하기 어색한 경우에 유용하다.

2.3 Speech Synthesis(SSML)

- Requirements-1999년 12월 23일
- Candidate Recommendation-2003년 12월 18일
- Recommendation-2004년 9월 8일

SSML은 미리 녹음된 음성의 조합, 음성과 음악의 합성 등을 통해 사용자에게 문장을 합성하는 방법을 제어할 수 있는 마크업 언어이다. 합성음의 특징(이름, 성별, 나이)과 속도, 볼륨, 피치, 강조 등을 표시 할 수 있다. 합성 엔진의 기본 합성음을 오버라이딩(overriding)하기 위한 규정 또한 존재한다.

2.4 Semantic Interpretation

- Last Call Working Draft-2004년 11월 8일
- Candidate Recommendation-2005년 말 예정

의미 해석 스펙은 인식 결과로부터 의미를 뽑아내기 위한 인식 문법에 대한 주석을 기술한다. 주석은 ECMAScript에 기반한 구문으로 표현 된다. 그리고 실행될 때, ECMAScript 변수 값으로 혹은 XML로 결과가 나타난다. XML 형태의 결과는 EMMA (Extensible Multimodal Annotation Markup Language)이다.

2.5 Call Control(CCXML)

- Requirements-2001년 4월 13일
- Last Call Working Draft-2005년 6월 29일
- Candidate Recommendation-2005년 말

CCXML은 음성 자원와 텔레포니 자원의 제어를 가능하게 하는 마크업이다. 이 언어 특징은 전화 교환 시스템에서 단말 교환 장치의 자원을 제어 하는 것이다. 전화망 음성 응용 개발자로 하여금 호 제어(call screening, call waiting, call transfer 등)를 할 수 있도록 한다. 사용자는 outbound call, 조건부 answering, 컨퍼런스 call 등 전화망 서비스 응용을 개발할 수 있다.

음성 브라우저에서 가장 핵심 표준인 대화 마크업 언어(VoiceXML)의 다음 버전(3.0) 목표는 진보된 음성정보시스템에서 사용될 수 있는 효과적인 대화 기술을 가능하게 하고, 다른 W3C 언어와 쉽고 명확하게 통합할 수 있는 형태를 제공하는 것이다. 예를 들어 멀티모달 인터페이스를 위해서는 차기 대화 마크업과 멀티모달을 구성하는 다른 모달리티를 위한 다양한 마크업 언어간의 통합이 가능해야 한다. VoiceXML 2.0와 비교하면, 개선된 대화 기능과 높은 융통성, 그리고 XHTML과 SMIL같은 언어에 내포될 수 있는 모듈화를 지원하게 될 것이다. 2003년 초, 차기 대화 마크업 언어를 위한 세부적인 요구 사항을 수집하는 것으로부터 작업은 시작되었다. 요구 사항들은 VoiceXML 2.0, Voice Browser group(특히 Call Control), SALT 1.0과 XHTML+Voice Profile 등과, W3C especially Multimodal, XHTML, WAI 같은 W3C 내의 다른 working group에서 제안되는 요구 사항들로 구성될 예정이다. 초안(first working draft)은 2005년 3분기에 공표될 예정이다.

다음은 VoiceXML로 기술된 음성 대화 시나리오의 한 예이다.

```
1. <?xml version="1.0" encoding="euc-kr"?>
2. <vxml version="2.0">
3. <form>
4.    <block>
5.       <prompt>
6.          <audio src="./SampleVXMLdoc/ music.wav"/>
7.       </prompt>
8.    </block>
9. </form>
10. <menu scope="document">
11.    <prompt>
12.    메뉴를 선택하십시요. <enumerate/>
13.    </prompt>
14.    <choice next="./SampleVXMLdoc/currentweather.vxml">
         현재날씨</choice>
15.    <choice next="./SampleVXMLdoc/
         tomorrowweather.vxml">내일날씨</choice>
16.    <choice next="./SampleVXMLdoc/weekweather.vxml">
         주간예보</choice>
17.    <choice next="./SampleVXMLdoc/worldweather.vxml">
         세계날씨</choice>.
18. </menu>
19. </vxml>
```

여기서, "music.wav"라는 이름의 파일에 미리 녹음된 소리가 사용자에게 재생된다(6번). 재생이 끝나면, "메뉴를 선택하십시요"라는 문장이 합성음으로 변환되어 사용자에게 프롬프트된다(12번). 사용자는 '현재날씨', '내일날씨', '주간예보', '세계날씨'의 4가지 중에 하나를 발화하여 다음 대화를 선택할 수 있다. 또한, 14번 줄에서 〈choice〉 마크업이기술된 순서에 따라, DTMF 키를 눌러 다음 대화를 선택할 수 있다. DTMF 키 '1'은 '현재 날씨', '2'는 '내일날씨', '3'은 '주간예보', '4'는 '세계날씨'에 해당한다. 사용자가 '3' 또는 '주간예보'라고 발화하면, 16번 줄의 〈choice〉 마크업에 next속성에 명시된 "weekweather.vxml" 파일

이 다음에 진행할 음성 대화 시나리오로 선택되어, VoiceXML 해석기가 이 파일을 해석하여 대화를 진행한다.

3. 음성기반 멀티모달 인터페이스 표준화

멀티모달 인터페이스는 2개 이상의 입력 모드를 동시에 사용하는 인터페이스 시스템을 의미한다. 정보시스템과의 인터페이스에서 멀티모달 인터페이스의 필요성은 크게 2 가지이다. 먼저, 단일 모달 입력의 신뢰도가 상황에 따라 많이 저하되는 데, 이를 다른 모달의 입력을 이용하여 보완하기 위해서이다. 예를 들면, 잡음이 많은 상황에서 음성 인식은 그 인식결과의 신뢰도가 매우 떨어진다. 이러한 한계를 극복하기 위하여, 음성 인식과 함께 제스춰 인식을 사용하여 사용자의 의도를 보다 정확히 인식할 수 있다. 다른 하나는 사람 사이의 상호작용은 기본적으로 멀티모달이며, 제스춰와 함께 지시대명사 등의 사용 등이 일반적이 상호작용 형태이기 때문에 일반 사용자가 가장 쉽게 사용할 수 있다는 점이다.

3.1 SALT(Speech Application Language Tags)

마이크로 소프트 사가 주도하는 SALT [6]는 HTML과 다른 마크업 언어(HTML, XHTML, WML)의 확장으로, 음성만을 위한 브라우저(VoiceXML과 동일한 목적)와 멀티모달(Multimodal) 브라우저의 2 가지 목표를 달성하기 위한 명세이다. SALT는 기본적으로 PC용 비주얼 웹 브라우저를 위한 HTML 또는 XHTML이나, 휴대전화 또는 PDA용 웹 브라우저를 위한 WML에 내포될 수 있도록 설계되었다. 따라서, SALT는 SALT는 비주얼 페이지를 통해 음성 입출력을 동시 지원할수 있는 멀티모달 인터페이스를 기술할 수 있는 언어 명세이다.

멀티모달 접근은 사용자가 여러 가지 방식으로 정보시스템과 상호 작용을 할 수 있도록 한다. 음성, 키보드, 키패드, 마우스 등을 이용해서 입력을 할 수 있고, 합성 음성, 오디오, 텍스트, 비디오, 그래픽 등과 같은 데이터를 산출할 수 있다. 또한, 비주얼 디스플레이가 없는 경우, SALT는 HTML 이벤트 모델과 스크립팅 모델을 사용하여 다이얼로그의 상호 작용 흐름을 관리하도록 하고 있다. 음성인식 문법은 W3C의 SRGS를 그대로 사용한다.

3.2 X+V(XHTML + Voice)

IBM이 주도하는 컨소시움에 의하여, 개발되고 있는 X+V [7]는 SALT와 매우 유사한 형태 및 목적을 가지고 지고 있다. SALT와의 차이는 SALT에서는 음성과 관련한 마크업을 VoiceXML과는 별도로 개발하였으나, X+V에서는 VocieXML을 그대로 사용한다.

이름에서도 나타나 듯이, X+V는 XHTML을 호스트 언어로 하여, VoiceXML을 내포 언어로 채용한 것이다. 또한, XML의 이벤트 모델을 그대로 사용할 수 있도록 하였다.

X+V와 SALT는 다음에 설명하는 W3C의 멀티모달 인터페이스 표준에 채택되기 위하여 제출된 상태로, 마이트로 소프트와 IBM 진영이 경쟁하는 모양을 보이고 있다.

3.3 W3C 멀티모달 인터페이스 표준화 활동

음성과 제스쳐로 접근할 수 있는 웹 페이지를 모토로 하여, W3C은 인터페이스의 다양한 모드를 지원하는 이동형 장치에서 다양한 인터페이스 모드를 지원하기 위한 표준의 개발을 목표로 표준화 활동을 진행하고 있다[8].

데스크탑 PC 시스템은 웹에 접근하기 위해 매우 효과적인 것이 증명되었다. 고해상도 화면, 포인팅 장치와 키보드는 많은 정보와 효율적으로 상호 작용하는 것을 쉽게 한다. 그러나 이동 중인 경우에, 포켓 또는 지갑에 꼭 맞는 작은 경량의 장치를 필요로 한다. 휴대전화는 매우 대중화된 장치이지만 표시할 수 있는 정보 양이 제한적인 표시 장치, 작은 수의 키 등의 제약이 있다. XHTML, CSS, SMIL 과 SVG 등의 W3C 권고안에서 이동 장치를 위한 모바일 프로파일이 포함되고 있으며, 이동 중에 웹의 접근은 현실화 되었다. 그러나 이동 장치의 작은 키패드만으로는 수 천 개의 문자, 표의언어를 위한 검색, 또는 웹 주소를 입력하는 것이 매우 힘들다. 최근 몇 년 사이에 웹을 전화로 접근하기 위한 수단으로 음성을 사용하는 것에 대한 관심이 집중 되었다. 그 결과로 W3C는 음성 인터페이스 프레임워크를 규정하고, 이에 필요한 표준안을 개발하였다.

VoiceXML 기반의 음성 인터페이스는 미리 녹음한 음성이나, 합성음을 이용하며, 단어와 간단한 구를 인식할 수 있는 단계에 까지 이르렀으며, 기술의 발정에 따라, 보다 자연스러운 대화체 인식 및 합성이 가능해질 것으로 기대되고 있다. 이에 따라, 음성 인터페이스를 다른 다양한 인터페이스 모드(특히 이동 가능한 휴대장치의 제한적인 입력 시스템 또는 펜, 터치 스크린 등)와 결합하고자 하는 연구가 촉진되고 있다. 음성과 다른 인터페이스 모드의 결합으로 나타나는 멀티모달 인터페이스는 사용자에게 말하고, 듣고, 쓰고, 타이

핑하고, 보는 보다 다양하고 자연스러운 시스템과의 상호 작용을 가능하게 할 것이다.

멀티모달 응용은 장치 성능, 사용자 모달리티 선호도와 환경 조건에 따라 대응할 수 있어야 한다. 예를 들면, 주위의 소음이 많은 곳에서는 음성 입출력을 비활성화 시켜야 한다. 그리고 다양한 사용 환경에 맞게 동적으로 대응할 수 있는 응용의 개발을 쉽게 할 수 있도록 지원하여야 한다.

3.3.1 Multimodal Interaction Framework

멀티모달 인터페이스 적용 시스템의 일반적인 구조, 와 구성 요소 및 사용할 수 있는 표준 및 마크업 언어와 관계를 명기하고 있다. 멀티모달 응용 사례 연구와 요구사항 분석을 다음과 같이 수행하였다.

- Multimodal Interaction Use Cases(2002년 12월 4일): 멀티모달 인터페이스의 이해를 돕기 위하여 멀티모달 인터페이스의 응용 예를 제시하고 설명함.
- Multimodal Interaction Requirements(2003년 1월 8일): 멀티모달 인터페이스 표준화를 위한 요구사항 분석서.

초안에는 입/출력 구성요소, 상호 장용 관리와 보조 구성요소에 초점을 맞출 예정이며, 다음으로 객체지향 모델기반의 마크업 개발과 다른 W3C 마크업 언어와의 통합 방법이 포함될 예정이다.

3.3.2 Extensible Multimodal Annotation Markup Language(EMMA)

입력 장치들과 멀티모달 상호작용 관리 시스템 사이의 인터페이스를 위한 데이터 형식을 표준화하기 위한 활동으로 다음과 같은 계획 하에 진행되고 있다.

- Requirements - 2003년 1월 13일
- Last Call Working Draft - 2005년 9월 16일

EMMA에는 인식기가 응용 별 특징 데이터와 부가적인 인식 스코어, 시간, 입력 모드, 그리고 부분의 인식 결과 등을 표기할 수 있는 방법을 정의할 예정이다. EMMA는 음성 브라우저 프레임워크 안에서 개발되는 의미 해석(semantic interpretation specification) 표준의 데이터 형식이기도 하다.

3.3.3 펜 입력(InkML)

여기서는 멀티모달 시스템의 전자펜이나 스타일러스에서 사용되는 잉크를 위한 XML 형식을 정의한다. 필기체 인식, 제스처 및 그림 인식과 수학, 음악, 화학 등에서 사용되는 특수 기호의 인식을한 것이다. IBM, Intel, Motorola, 그리고 International Unipen Foundation에서 진행하여 온 연구를 바탕으로 출발하였으며, 다음과 같은 계획을 가지고 있다.

- Requirements- 2003년 1월22일
- Last Call Working Draft - 2005년 4월

4. MRCP(Media Resource Control Protocol)

MRCP[10]는 IETF(Internet Engineering Task Force)의 Speech Control Working Group에 의하여 진행되고 있는 분산 음성 처리 및 응용을 위한 통신 프로토콜이다. 그림 3은 MRCP의 구조를 보이고 있다. MRCP는 음성 인식/합성/인증 등 고성능 하드웨어를 필요로 하는 음성 처리 모듈은 서버에 두고, 작은 용량의 이동 기기에서 음성 서버에서 제공하는 인식/합성/인증 서비스를 IP 기반의 프로토콜을 통하여 지원받을 수 있도록 한다.

그림 3에서와 같이, MRCP는 음성 인식/합성/인증 서비스의 요청과 제어를 위한 통신 프로토콜이며, RTP나 SIP와 같은 실시간 스트리밍 프로토콜과 함께 구성된다. 예를 들면, 특정한 문장의 합성이 필요하면, MRCP를 이용하여 클라이언트가 음성 서버에 합성을 위한 문장을 전달하고 합성을 요구한다. 서버에서 합성된 음성 신호는 RTP나 SIP와 같은 실시간 스트리밍 프로토콜을 이용하여 클라이언트에 전달하여, 사용자에게 합성음을 들여 준다.

통신 프로토콜이지만 MRCP 자체가 음성 정보 처리를 위한 표준 API이기 때문에 MRCP 클라이언트는 다수의 이기종 음성 서버에 수정없이 접속할 수 있다. MRCP 클라이언트와 서버는 특정 컴퓨터와 운영체제에 독립적으로 구현할 수 있어, 플랫폼 독립적인 구성이 가능하다. 이는 유비쿼터스 컴퓨팅에서 매우 중요한 요소이다.

MRCP에서 클라이언트와 서버가 주고 받는 메시지는 앞에서 언급한, SRGS(Speech Recognition Grammar Specification), SSML(Speech Synthesis Markup Language) 등을 이용하도록 설계되어 있다.

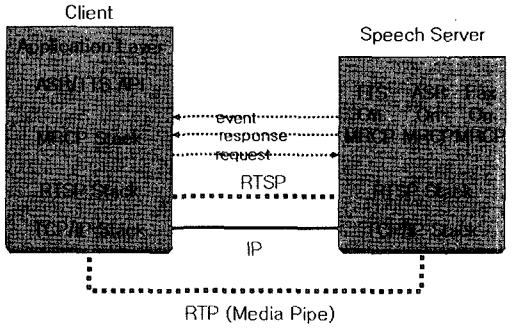

그림 4 MRCP 구조

5. 결 론

본 고에서는 유비쿼터스 환경에서 음성정보기술의 국제 표준화 동향에 대하여 살펴보았다. 음성정보시스템 개발을 위한 표준은 MS SAPI, JSAPI 등 함수 수준의 표준 명세에서 VoiceXML, SALT 등 웹 기반의 개방형 표준으로 발전하였다. 웹 기반의 개방형 표준은 음성 대화, 음성 인식/합성, 정보시스템, 전화 망 등의 접속 망을 상호 분리하여, 음성 정보시스템 구성 요소인 이들의 상호 독립적인 개발을 보장한다. 또한, 표준을 따르는 음성 브라우저의 개발로 음성 인식/합성이나 전화 망 등에 대한 이해가 없이도 음성 대화만을 설계 기술하여 음성정보시스템을 개발할 수 있도록 함으로써 음성정보기술의 보급 및 확산에 크게 기여하고 있다.

현재, W3C에서는 음성 브라우저에서 더 나아가 음성을 중심으로 한 멀티모달 인터페이스 개발을 위한 개방형 표준을 개발 중이며, MS SALT와 VoiceXML을 XHTML(HTML의 XML 버전)에 내포한 X+V [7]가 멀티모달 인터페이스 표준 명세로 제안되어 있다. 실제로, 국외에서는 VoiceXML 및 SALT를 기반으로 하는 음성정보 응용이 매우 빠르게 확산되고 있다. 유비쿼터스 환경과 같이 분산 환경이면서, 다수의 이기종 클라이언트, 서버를 활용해야 하는 경우의 음성정보 시스템을 위하여, MRCP와 같은 음성 서비스 제어를 위한 통신 프로토콜 표준이 진행 중이다.

국내에서는 이러한 개방형 표준을 기반으로 하는 음성정보시스템의 개발이 아직 미진한 상태이며, VoiceXML 등 이미 국외에서는 상당한 보급과 확산이 이루어진 표준에 대해서도 국내에서 구현된 시스템이 드문 상황이다. 키 보드, 마우스, 대형 모니터 등과 같은 기존의 사용자 인터페이스의 사용이 원활하지 못한 유비쿼터스 환경에서 음성은 매우 중요한 사용자 인터페이스로 생각되고 있으며, 이를 위한 기술 표준의 개발이 국외에서는 활발히 진행되고 있다. 국내에서도 음성 정보 기술의 표준화에 적극 참여하여, 국내 음성 정보 처리 기술 및 응용 개발의 국제 경쟁력을 높이기 위한 노력이 필요한 때이다.

참고문헌

[1] Java Speech API Home, http://java.sun.com/products/java-media/speech May, 2003.

[2] The Java Telephony API: an Overview, http://java.sun.com/products/jtapi/jtapi-1.2/Overview.html, Oct., 1997.

[3] Microsoft Windows CE .Net 4.2 SAPI 5.0 Overview, http://msdn.microsoft.com/library/default.asp?url=/library/en-us/wcesapi/html/ceconsapi50overview.asp, April, 2004.

[4] Voice Browser Activity Voice enabling the Web, http://www.w3.org/Voice, Dec. 2005.

[5] Voice Extensible Markup Language(VoiceXML) Version 2.0, http://www.w3.org/TR/2003/CR-voicexml20-20030128, W3C Candidate Recommendation, January, 2003.

[6] Speech Application Language Tags(SALT) 1.0 Specification, http://www.saltforum.org/saltforum/downloads/SALT1.0.pdf, July, 2002.

[7] X+V 1.1 XHTML+Voice Profile, http://www.voicexml.org/specs/multimodal/x+v/11, May 2003.

[8] Multimodal Interaction Activity, http://www.w3.org/2002/mmi, Feb. 2004.

[9] David Pears, "Enabling New Speech Driven Services for Mobile Devices: An overview of ETSI standard activities for Distributed Speech Recognition Front-ends," AVIOS 2000: The Speech Applications Conference, May 2000.

[10] Saravanan Shanmugham, et al., Media Resource Control Protocol Version 2(MRCPv2), internet draft of IETF(Internet Engin- eering Task Force) Speech Control Working Group, May 2004.

홍 기 형

1985. 2 서울대학교 컴퓨터공학과(학사)
1987. 2 한국과학기술원 전산학과(석사)
1994. 2 한국과학기술원 전산학과(박사)
1994. 2~1998. 2 한국전자통신연구원 데이터베이스팀 선임연구원
1998. 3~현재 성신여자대학교 미디어 정보학부 부교수
관심분야 : 음성응용, XML, 멀티모달 인터페이스, 멀티미디어 DB
E-mail : khhong@sungshin.ac.kr

정 민 화

1984. 2 서울대학교 제어계측공학과(학사)
1988. 2 Univ. of Southern California 전기공학과(석사)
1993. 8 Univ. of Southern California 전기공학과(박사)
1993. 12~1994. 7 한국통신 연구개발원 선임연구원
1994. 9~2004. 7 서강대학교 컴퓨터학과 부교수
2004. 8~현재 서울대학교 언어학과 부교수
관심분야 : 음성언어처리, 음성인식, 언어모델
E-mail : mchung@snu.ac.kr