

특 집

음성정보 기술 응용 개발 표준화

흥 기 형

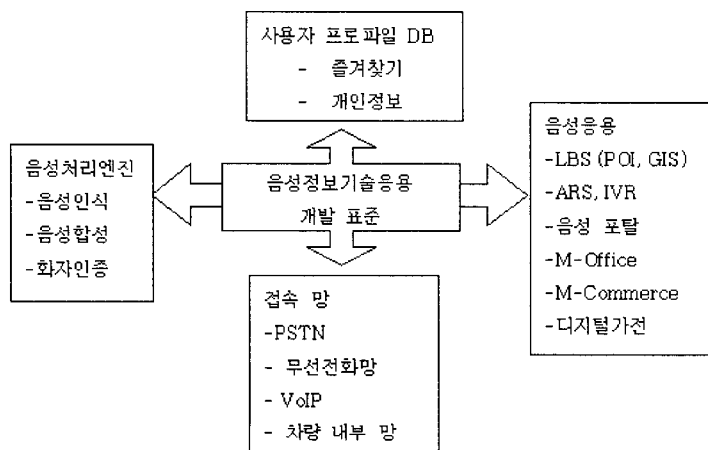
성신여자대학교 미디어정보학부

I. 서 론

음성인식, 합성, 화자인증 등의 음성정보 원천 기술 수준이 상용화 단계에 접어들고, 이동 중과 같이 모니터 등 시각적 인터페이스의 사용이 용이하지 않은 상황에서도 정보 접근의 요구가 증대함에 따라 음성은 정보시스템의 중요한 인터페이스로 자리매김하고 있다. 기존의 ARS(Automatic Response System), IVR(Interactive Voice Response) 시스템에서, 자동차용 텔레매틱스, 디지털 가전 및 홈 오토메이션 등으로 음성정보기술 응용 분야가 확대되고 있으며, 이에 따라 정보기술을 다양한 응용에 적용하기 위한 편리한 응용 개발 환경과 음성 응용 프로그램의 플랫폼 독립성 및 호환성의 확보가 절실히 요구되고 있다.

음성정보기술 응용 표준은 <그림 1>에서 보인 바와 같이 음성정보시스템을 구성하는 4가지 시스템의 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.

음성처리엔진은 음성정보 원천기술에 해당하며, 접속 망은 사용자가 음성정보에 접근하기 위하여 사용하는 망으로 과거에는 공중전화 유선 망(PSTN)이 유일하였으나, 최근에는 VoIP(Voice over IP)를 기반으로 하는 인터넷 망, 휴대전화와 같은 무선 음성 및 데이터 망, 차량 내부의 블루투스, CAN(Car Area network)과 같은 단거리 무선 망 등 매우 다양해지고 있



<그림 1> 음성정보응용기술 표준과 응용 및 개발 환경

다. 음성응용은 기존 ARS, IVR 뿐아니라 LBS (Location-based Service), 모바일오피스(M-Office), 모바일커머스(M-Commerce) 등으로 확대되고 있다.

음성정보기술 응용 표준은 각기 다른 기관이 개발한 음성 인식/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 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.

본 고에서는 음성정보기술 응용 개발과 관련된 국제 표준에 대하여 알아보려고 한다. 먼저, II장에서는 기존 음성 응용 개발의 문제점을 살펴 보고, 웹 기반 개발 환경의 장점에 대하여 간략히

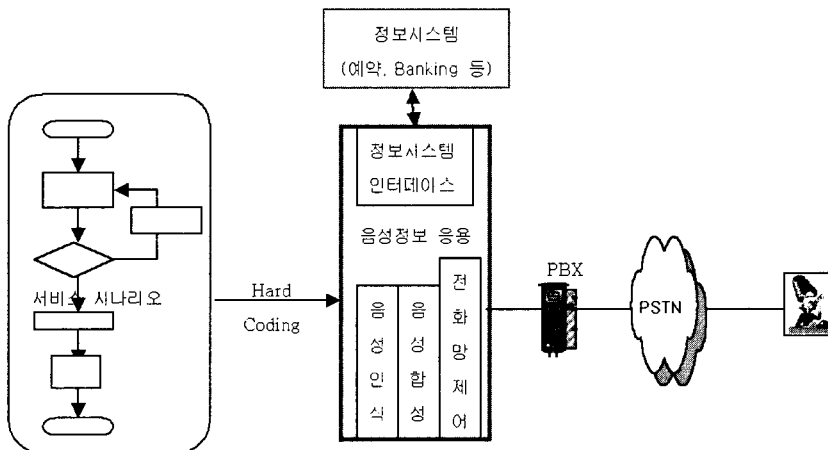
기술하였다. III장에서는, 현재 가장 급속하게 성장하고 있는 음성정보기술 응용 분야인 인터넷 정보의 음성을 통한 브라우저를 위하여, 그리고 웹 개발 환경을 음성응용 개발에 적용하기 위한 W3C(World Wide Web Consortium)의 음성 브라우저(Voice Browser) 관련 표준을 알아보았다. IV장에서는 정보시스템 플랫폼에서 전세계 시장의 상당 부분을 차지하고 있는 마이크로소프트사가 주축이 되어 제안하고 있는 SALT(Speech Application Language Tags)를 알아보고, 마지막 결론으로 향후 음성 응용 개발 표준의 방향에 대하여 기술하고자 한다.

II. 기존 음성정보기술 응용 개발 환경과 웹

대체로 음성정보기술 응용 시스템은 다음과 같이 4개의 모듈로 나눌 수 있다.

- 음성 대화 시나리오
- 정보시스템과의 인터페이스
- 전화망 인터페이스
- 음성인식/합성 엔진 인터페이스

90년대 이전의 음성정보기술 응용은, PSTN을 통한 전화환경에 맞추어 개발되었으며, 현재



〈그림 2〉 기존 음성정보 응용 개발 및 운영 환경

도 국내의 대부분의 음성응용 (텔레뱅킹, 콜센터 ARS, 철도청 ARS 예약 시스템 등)이 이러한 환경에서 개발되어 있다. <그림 2>는 이러한 기존 시스템의 구성을 보이고 있다.

<그림 2>에서 각종 예약 시스템이나 은행의 정보 시스템을 위한 음성 인터페이스를 개발하는 과정은 먼저, 음성으로 서비스할 시나리오를 설계하고, 설계한 시나리오를 하드 코딩(C나 C++ 등의 언어로)하여 음성정보 응용 프로그램으로 구현한다. 음성정보 응용 프로그램에는 서비스 시나리오 뿐 아니라 서비스 시나리오 중에 필요한 전화망 제어(고객으로부터의 전화 받기, 전화 끊기, DTMF 인식 등)와 예약 및 뱅킹 시스템과의 데이터 통신 등 정보시스템과의 인터페이스가 함께 프로그래밍 된다. 음성 인식/합성 엔진을 사용하는 경우에는 음성 인식/합성 엔진과의 인터페이스까지 하나의 프로그램으로 구현된다.

이러한 기존 방식의 문제점은 다음과 같다.

- 각 모듈 사이의 종속성 : 음성정보 응용 프로그램에 음성 서비스를 위한 음성 대화 순서, 정보 시스템과의 인터페이스, 전화망 제어, 인식/합성 엔진과의 인터페이스가 함께 프로그래밍 되어 있어, 각 모듈 간의 독립성이 없다. 이에 따라, 교환기의 교체 등으로 인한 링 톤(ring tone) 등의 변화, 보다 성능 좋은 음성 인식/합성 엔진로의 교체, 정보 시스템의 변경이 있을 때, 음성 정보 응용 프로그램 전체를 수정해야 하고, 새로 코딩을 하여야 한다.
- 서비스 변경시 고비용 : 만일 음성응용의 시나리오, 즉 대화 순서의 변경이나 추가가 필요할 경우, 기존 음성 정보 응용 프로그램 전체를 재개발해야 한다. 프로그램의 수정, 컴파일, 시험 등의 과정을 그대로 거쳐야 한다.
- 음성 응용 개발자의 고도 전문성 요구 : 하나의 프로그램에 음성정보 서비스를 위한 음성 대화 시나리오, 음성 인식/합성 엔진 인터페이스, 전화망 제어/정보시스템과의 인터페이스가 결합되어 있으므로, 음성정보 응용 프로그램 개발자는 이들 모두에 대한 이해와 전문성을 가지

고 있어야 한다.

- 맞춤형 서비스 제공 불가 : 하나의 시나리오가 하드 코딩된 상태이므로 고객의 특성에 따른 고객별 맞춤형 대화 서비스의 제공을 할 수 없다.

이외에도, 시스템이 대화의 주도권을 가지는 음성 대화 시나리오로 사용자 (사람)에게 매우 불편한 느낌을 가지도록 하는 등의 단점이 존재한다.

모듈 사이의 종속성을 해결하기 위하여, Java의 경우, JSAPI(Java Speech API)^[1], JTAPI(Java Telephony API)^[2], 마이크로 소프트웨어사의 MS SAPI(Speech API)^[3] 등이 개발되었다. 이들 API 수준의 산업체 표준은 음성 인식/합성기와 전화망 제어 기능을 각 사의 표준화된 API로 분리함으로써, 음성 응용 개발자는 이들 API만을 숙지하면, 이를 이용하여 음성 대화 시나리오를 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 사용자

의 모니터에 실현해준다. 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 다른 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 정보시스템 사이트로 접근하여, 링크되어 있는 다른 정보 시스템에 쉽게 접근할 수 있다.

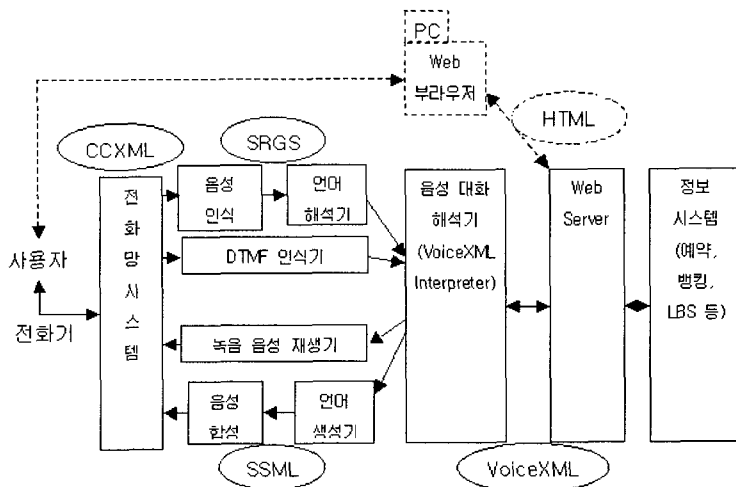
상기의 장점으로 인하여, 현재 웹은 거의 모든 정보 시스템의 인터페이스 개발 및 운영 환경으로 자리 잡고 있다. 그래픽 모니터의 특성이나, 정보 시스템과의 인터페이스에 대한 이해가 없어도 간단한 HTML 마크업 만을 알면, 초등학교

이라도 작성할 수 있으므로, 이에 따라 HTML의 확산이 매우 빠르게 진행되었다.

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

III. W3C 음성 브라우저 표준화

<그림 3>은 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



<그림 3> W3C 음성 브라우저

대화(dialog)를 기술하기 위한 음성 대화 기술 마크업 (dialog markup) 언어이다. 이러한 XML 기반의 음성 대화 및 인식/합성 마크업은 웹 기반 콘텐츠의 음성 서비스를 가능하게 할 뿐 아니라, II장에서 기술한 웹 기반 개발의 장점을 음성정보 응용 개발에서 가능하도록 한다.

1. VoiceXML

VoiceXML^[5]은 합성된 음성, 디지털화된 오디오, 음성과 DTMF 키 입력의 인식과, 음성 녹음, 전화통신, 상호 주도 대화 등의 특징이 있는 음성 대화 기술 언어이다. 목적은 대화형 음성언어 인터페이스를 위한 음성 입출력에 기반한 표준화된 음성 시나리오를 제공하고, 하위-레벨 프로그래밍과 리소스 관리로부터 응용, 즉 대화 시나리오의 개발과 운영을 분리하여 독립시키기 위한 것이다. VoiceXML을 사용함으로써 가질 수 있는 장점은 다음과 같다.

- 다양한 서비스 및 콘텐츠 제공자가 나타날 수 있다. 즉, 정보 제공자는 VoiceXML 표준에 따라 음성 입출력이 가능한 시나리오를 만들 수 있고, 장비 제공자는 서비스 시나리오에 독

립되게 장비를 구축할 수 있으므로 다양한 음성 서비스가 신속히 개발될 수 있다.

- 기존의 웹 기반기술을 활용할 수 있다.
- 클라이언트와 서버의 상호작용을 최소화할 수 있다

다음 <그림 4>는 VoiceXML로 기술된 음성 대화 시나리오의 한 예이다.

<그림 4>의 예는 먼저, “music.wav”라는 이름의 파일에 미리 녹음된 소리가 사용자에게 재생된다(6번). 재생이 끝나면, “메뉴를 선택하십시오”라는 문장이 합성음으로 변환되어 사용자에게 프롬프트된다(12번). 사용자는 ‘현재날씨’, ‘내일 날씨’, ‘주간예보’, ‘세계날씨’의 4가지 중에 하나를 발화하여 다음 대화를 선택할 수 있다. 또한, 14번 줄에서 <choice> 마크업이 기술된 순서에 따라, DTMF 키를 눌러 다음 대화를 선택할 수 있다. DTMF 키 ‘1’은 ‘현재 날씨’, ‘2’는 ‘내일 날씨’, ‘3’은 ‘주간예보’, ‘4’는 ‘세계날씨’에 해당한다. 사용자가 ‘3’ 또는 ‘주간예보’라고 발화하면, 16번 줄의 <choice> 마크업에 next 속성에 명시된 “weekweather.vxml” 파일이 다음에 진행

```

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 선택하십시오. <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>

```

<그림 4> VoiceXML의 예제

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

2. Speech Recognition Grammar(SRGS)

SRGS는 음성과 DTMF 입력 둘 다를 포함하며, 음성 인식이나 DTMF 인식 문법을 기술하는 표준 명세이다. DTMF는 시끄러운 상태나 문맥상 말하기 어색한 경우에 유용하다. 음성 인식은 본래 불확실한 프로세서이다. 어떤 음성 엔진은 “um’s”과 “aah’s”를 무시하거나, 부분적인 매치를 수행시키는 것이 가능하다. 인식기는 정확한 값을 통보해야 한다. 가능한 인식 단어 후보가 다수 존재하면, 인식기는 가장 최적의 인식 단어를 골라 낼 수 있어야 한다(n-best results).

3. Speech Synthesis(SSML)

SSML은 미리 녹음된 음성의 조합, 음성과 음악의 합성 등을 통해 사용자에게 문장을 합성하

는 방법을 제어할 수 있는 마크업 언어이다. 합성 음의 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.

4. Call Control(CCXML)

CCXML은 음성 자원과 텔레포니(전화 망) 자원의 제어를 가능하게 하는 마크업이다. 이 언어 특징은 전화 교환 시스템에서 단말 교환 장치의 자원을 제어 하는 것이다. 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 등을 XML 형태로 기술할 수 있다.

다음 <표 1>은 W3C의 음성 브라우저 활동에서 진행중인 다양한 음성응용 개발 관련 표준 명세의 종류와 현 상태를 보이고 있다.

<표 1> W3C 음성 브라우저 표준화 현황 (2003년 5월 현재)

분 류	표 준	현재 상태
VoiceXML	Voice dialog requirements	WD, 23 December 1999
	VoiceXML 2.0	CR, 18 Jan. 2003
Speech Synthesis	Speech synthesis requirements	WD, 23 December 1999
	Speech synthesis specification	WD, 2 December 2002 (Last call)
Speech Recognition	Speech grammar requirements	WD, 23 December 1999
	Natural language processing requirements	WD, 23 December 1999
	Speech Grammar specification	CR, 26 Jun 2002
	Semantic Interpretation for Speech Recognition	WD, 1 April 2003
	Stochastic language models	WD, 3 January 2001 (1st draft)
	Natural language semantics markup language	WD, 20 November 2000 (1st draft)
Pronunciation Lexicon	Pronunciation lexicon requirements	WD, 12 March 2001
Call Control	Call control requirements	WD, 13 April 2001
	Call Control XML (CCXML)	WD, 11 October 2002

IV. SALT(Speech Application Language Tags)

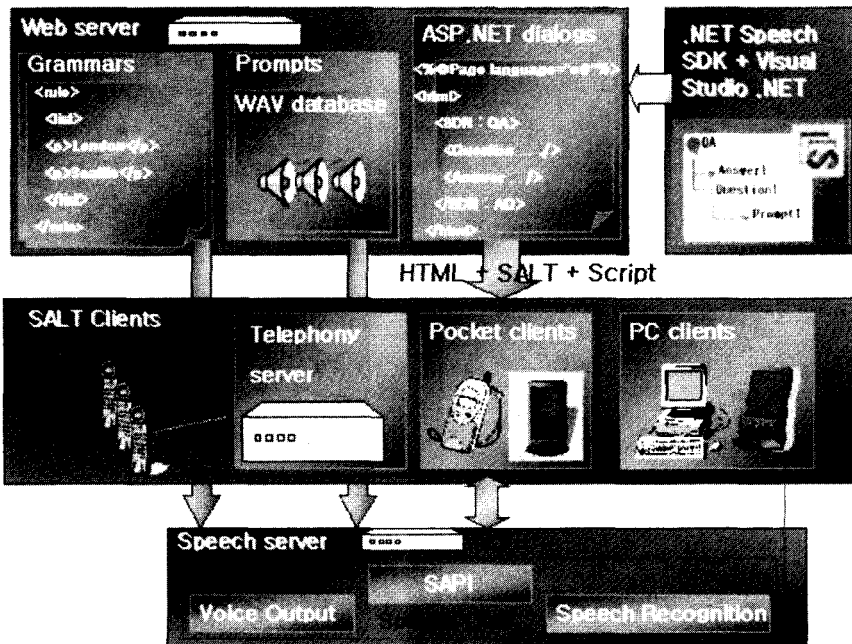
마이크로 소프트 사가 주도하는 SALT^[6]는 HTML과 다른 마크업 언어(HTML, XHTML, WML)의 확장으로, 음성만을 위한 브라우저 (VoiceXML과 동일한 목적)와 멀티모달 (Multimodal) 브라우저의 2가지 목표를 달성하기 위한 명세이다. SALT는 기본적으로 PC용 비주얼 웹 브라우저를 위한 HTML 또는 XHTML이나, 휴대전화 또는 PDA용 웹 브라우저를 위한 WML에 내포될 수 있도록 설계되었다. 따라서, SALT는 비주얼 페이지를 통해 음성 입출력을 동시 지원할수 있는 멀티모달 인터페이스를 기술할 수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

멀티모달 접근은 사용자가 여러 가지 방식으로 정보 시스템과 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 있

다. 또한, 비주얼 디스플레이가 없는 경우, SALT는 HTML 이벤트 모델과 스크립팅 모델을 사용하여 다이얼로그의 상호 작용 흐름을 관리하도록 하고 있다.

SALT의 구현을 위해서는 <그림 5>에서와 같이 4가지 주요 구성 요소가 있다.

- ① 웹 서버 : 웹 서버는 HTML, SALT, 중첩된 스크립트를 포함하는 웹 페이지를 생성한다.
- ② 텔레포니 서버 : 텔레포니 서버는 전화 망과 연결한다. 이 서버는 HTML, SALT, 그리고 스크립트를 인터프리트하는 음성 브라우저를 포함한다. 각 고객 (caller)에 대해 개별 프로세스와 쓰레드에서 브라우저를 수행할 수 있다. HTML은 GUI를 기술하는 것으로 음성 브라우저와는 관련이 없기 때문에, 음성 브라우저는 단지 HTML에 포함되어 있는 SALT 마크업만을 해석한다.
- ③ 음성 서버 : 음성 서버는 음성을 인식하고, 오디오 프롬프트를 재생하거나 합성음을 생성하여 SALT 클라이언트 장치에게 전달한다.



<그림 5> SALT 구성 요소

④ 클라이언트 장치 : 예를 들자면, 클라이언트는 HTML과 SALT를 인터프리팅할 수 있는 인터넷 익스플로러 버전을 탑재한 포켓 PC, 데스크탑 PC가 될 수 있다. 또한, 표시장치가 없는 전화 사용자를 위해서는 텔레포니 서버가 SALT를 해석하여 사용자와 대화한다.

V. 결 론

본 고에서는 음성정보 응용 개발을 위한 표준에 대하여 살펴보았다. 음성정보 응용 개발 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, 전화망 등의 접속 망을 분리하였다. 표준을 따르는 음성 브라우저의 개발로 음성 인식/합성이나 전화망 등에 대한 이해가 없이도 음성 대화만을 설계 기술하여 음성 정보 응용을 개발할 수 있도록 한다.

실제로, 국외에서는 VoiceXML 및 SALT를 기반으로 하는 음성정보 응용이 매우 빠르게 확산되고 있다. 그러나, 국내에서는 이러한 개방형 표준을 기반으로 하는 응용의 개발이 아직 미진한 상태이며, 특히, 우리 말 음성 대화의 특징, 음성 합성에 있어서 우리 말이 가지는 특징을 고려한 표준 명세의 개발이 필요하다고 하겠다.

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.

“본고는 산업자원부 지원 중기거점 자동차용 음성 HMI 시스템 기술 개발 사업에서 일부 지원을 받았음.”

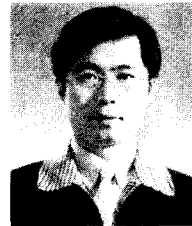
참 고 문 헌

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

[sun.com/products/java-media/speech](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, 2003.
- [4] Voice Browser Activity-Voice enabling the Web, <http://www.w3.org/Voice>, April 2003.
- [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.

저 자 소 개



홍 기 형

1985년 2월 서울대학교 컴퓨터공학과 졸업, 1987년 2월 한국과학기술원 전산학과 졸업(석사), 1994년 2월 한국과학기술원 전산학과 졸업(박사), 1994년 2월~1998년 2월 : 한국전자통신연구원 데이터베이스팀 선임연구원, 1998년 3월~현재 : 성신여자대학교 미디어정보학부 조교수, <주관심 분야> : 음성응용, XML, 멀티미디어 DB.>