음성 인터페이스의 기술 현황과 표준화 동향※

장민석* · 김성국**

1. 서 론

음성 인터페이스는 시간과 장소 및 장비에 구애받지 않는 유비쿼터스 환경에서의 편리한 사용자 인터페이스로 중요하게 부각되고 있다. 최근 다보스 포럼, MIT대학, 가트너 그룹, 뉴욕타임즈 등의 미래 예측에서 21세기 정보화 사회를 선도할 기술로 음성언어기술이 선정된 바 있다. 국내에서는 정보통신부 주관의 IT839 정책(8대 신규서비스, 3대 첨단 인프라, 9대 신성장 동력)으로 인해 텔레매틱스, 지능형로봇, 디지털홈, 차세대 PC, USN, U-City, DMB, HSDPA, WIBRO 등과 같은 정보기술의 혁신적인 발전을 도모하고 있으며, 이를 기반으로 하는 각종 정보기기가 소형화되고 유선 환경에서 모바일 환경으로 이동성이 강화되고 있다. 아울러 기존의 유·무선 전화망 환경에서 초고속 무선 데이터망으로 통신환경 진화 현상으로 인해 음성 인터페이스의 요구는 향후 폭발적으로 증가될 전망이다. 이에 따라 IT839의 HCI 분야의 주요 정책목표는 2012년까지 유비쿼터스 환경에서의 HCI 분야 소프트웨어 세계시장 선점이다. 이를 위해 음성언어기술 보급 및 인식확산 추진, 세계시장 선점을 위한 음성언어분야 전략기술 집중육성, 유비쿼터스 환경에서의 HCI 관련 법·제도 개선에 힘을 쏟을 예정이다.

하지만 음성이 인간의 가장 자연스럽고 편리한 정보교류의 수단임에도 불구하고 1950년대 이후 지속적으로 진행되어 온 음성 인터페이스 기술은 인간에게 아직은 만족스러운 단계에 이르지 못하고 있다. 음성 인터페이스의 궁극적인 목표가 인간과 마찬가지로 인간의 음성과 감정을 자유롭게 인식하고 발성하는 수준이라면 더더욱 요원한 단계에 머무르고 있다.

본고의 구성은 다음과 같다. 2장에서는 음성 인터페이스의 기술적인 전망을 살펴보며, 3장에서는 현재의 음성 인터페이스의 개발 현황 및 구현 유형을 살펴본다. 4장에서는 환경변화에 따른 표준화 동향을 알아보고 마지막으로 결론을 도출하고자 한다.

2. 음성 인터페이스의 기술적인 전망

앞으로의 음성 인터페이스 기술 및 환경변화 등을 전망해보면 다음과 같다[1].

환경변화 측면에서 보면 기존의 음성인식 서비스가 CTI(Computer Telephony Integration) 기반의 음성통신망을 사용하던 것이 향후 무선 데이터통신 속도의 향상에 의해 DSR(Distributed Speech Recognition) 형태의 음성인식 구조로 진

* 군산대학교 컴퓨터정보과학과 부교수
** 군산대학교 컴퓨터정보과학과 이학석사 과정중
※ 본 결과물은 정보통신부의 정보통신기초기술연구지원사업(정보통신연구진흥원)으로 수행한 연구결과입니다.

표 1. 음성인터페이스 기술 전망

변화 측면	분야	2004년 이전	2005년	2006년	2007년	2008년	2009~2012년
환경 변화	통신인프라	유·무선 전화망 기반 CTI		고속 무선데이터망(WIBRO)기반 DSR(분산음성인식)			
	사용자 인터페이스	단일 모덜리티 지원 (예:음성, 키보드/키패드, 마우스, 펜)		멀티 모덜리티로 다중 입력 (예:음성+키보드+펜+마우스)			멀티모덜 확장 (예:제스처, 비디오)
기술 변화	음성인식	중규모(수천~수만 단어급) 명령형 단어/핵심어인식		대규모(수십만~수백만)대화형 핵심어/연속 음성인식			감정인식
응용 변화	지능형 로봇/ 홈네트워크	제한영역 근거리 명령어 인식		제한영역 멀티모덜 기반 원거리 대화형 인식, 수천 단어 급의 핵심어 검출이 가능한 지능형 로봇		자유영역 멀티모덜 기반 원거리 대화형 감정인식	
	텔레매틱스	소규모 제어명령		멀티모덜 기반 대규모(수십만 단어급)목적지 인식		멀티모덜 대화형 정보검색	
기타 응용변화		명령어 중심, VAD, 자동 번역, 정보 제공 서비스, 예약 등	무인콜센터, ARS, IVR, 시스템 주도형 텔레매틱스 서비스	사용자 주도형 텔레매틱스 서비스, 대화형 음성인식, 웹 자동통역	다국어 인터넷 번역, 디지털 컨텐츠 자동 색인, 웹질의 응답 정보검색, 114 무인자동 안내, 개인통신분야(음성으로 SMS, MMS, e-mail, 메신저, 채팅 기능), 사용자 주도형 이동환경 멀티모덜 대화형 정보검색		음성인식 자동 A/S 서비스
사용자변화		명령(수동적)		대화(협동적)			공감(능동적)

화할 전망이다. 데이터 통신망 기반에서 음성인식이 이루어질 경우 다른 모덜리티(Modality)와의 융합이 용이해지므로 다양한 멀티모덜 인터페이스(MultiModal Interface)가 발전할 것으로 전망된다. 기술변화 측면에서는 인식대상 단어의 규모가 증가하며, 기존의 단어/핵심어인식 단계에서 핵심어/연속음성인식 더 나아가 감정인식으로 발전함으로써 인간 수준에 점점 다가갈 것으로 전망한다. 응용측면에서는 지능형 로봇/홈네트워크, 텔레매틱스, 차세대 PC 등 신성장동력 산업에의 응용분야가 점차 확대되고 사용자 편의성도 증대할 전망이다.

하지만 무엇보다 이러한 음성 인터페이스의 발전이 가능하기 위해서는 선결해야 할 문제가 있다. 기존의 음성 인터페이스의 환경에 비해 정보 환경의 이동성 및 유비쿼터스 환경 증가로 인한 잡음처리 연구 잡음처리 연구, 멀티모덜 인터페이스, 연속음성 인식, 대규모 인식, 분산 음성 인식(DSR), 감정 인식 기술, 임베디드 음성 인식 기술이 현재 활발히 연구 중에 있다. 참고로 멀티모덜 인터페이스는 음성, 마우스, 그래픽, 키보드/키패드, 펜(스타일러스) 등 다양한 입출력 방법을 동시에 복합적으로 사용하는 인터페이스 기술을 의미한다. 미래에는 촉각 장갑이나 인체 부착 센서와 같은 촉각 수단이 새로운 모덜리티가 될 전망이다[2].

3. 음성 인터페이스의 개발 현황 및 구현 유형

현재 다양한 형태의 제품(응용, 도구, 음성인식/합성 엔진, 서비스/플랫폼 등)에서 음성 인터페

표 2. 음성 인터페이스를 채택한 개발 예

응용 분야	제품명	회사명	채택 기술	표준지원 여부	비고
ARS, IVR	철도 예약 시스템	철도청	PSTN, 서버기반	-	ARS(Automatic Response System)
	MegaMPS	브리지텍	PSTN, 서버기반	-	폰뱅킹 IVR(Interactive Voice Response) 시스템 계좌번호, 주민번호 등을 음성으로 입력할 수 있는 서비스
CTI 기반 무인 콜센터	HMIHY	AT&T	PSTN, 서버기반	VoiceXML	"How May I Help You?" 서비스 자사 고객센터를 위한 콜센터 서비스
보이스 포털	Telsima Voice Portal on VPRS	HP + Intel	PSTN, 서버기반	VoiceXML	VPRS(Voice Portal Recognition System) France Telecom and Orange에 적용 뱅킹, 뉴스, PIM 등등 제공 Philsoft v3 인식엔진 AT&T Natural Voice; TTS 엔진 탑재
항공예약/예매, 전화망대화 연속음성인식, 방송뉴스 인식	-	미국의 DARPA 프로젝트 연구결과	PSTN, 서버기반	VoiceXML	출발지, 목적지 등을 발성하면 이를 인식하여 항공편을 조회하거나 예약할 수 있는 서비스
딕테이션 소프트웨어, 발음교정	Dragon Naturally Speaking	Scansoft	클라이언트 방식	-	엑셀, 워드, 아웃룩 익스프레스, 인터넷 IE 등에서 사용, 개발자용은 클라이언트와 서버용으로 구분
	영어발음교정 서비스	IBM	클라이언트 방식	-	퍼스널은 15만 단어, 표준형은 30만 단어 수록
	Dr.Speaking	보이스웨어	인터넷망, 클라이언트 방식	-	보이스웨어와 언어과학이 협력하여 출시한 제품으로 발음교정 프로그램
텔레매틱스 서비스 (차량 네비게이션, 위치추적, 인터넷 접속, 원격 차량진단, 사고 감지, 교통 정보 제공, 홈 네트워크 및 사무자동화와 연계하여 정보제공, 엔터테인먼트 등의 기능을 제공하는 서비스[주로 길안내, 교통정보 서비스 위주])	단말기 (eXride)	현대모비스	무선 데이터 통신, 클라이언트 방식	-	보이스웨어의 50단어 수준의 인식 엔진 채택 사용화(명령어+주소록, VAD)
	단말기 (MTSII, I)	LG 전자	"	-	보이스웨어의 음성인식 엔진 채택 (개발중, 명령어+주소록, 목적지 설정 인식 수준)
	네이트드라이브 서비스	SK텔레콤	무선 데이터 통신, 휴대폰, 서버기반	-	2002년 3월부터 마국 스피치웍스의 음성인식 솔루션과 국내 코아보이스의 음성합성 솔루션을 채용하여 서비스중 (인식범위; 길안내) 최초의 음성인식 길안내 서비스
	MOZEN 서비스	현대차	무선 데이터 통신, 텔레매틱스 단말기, 무선전화 음성합성;클라이언트 방식 인식은 서버기반	-	보이스웨어의 음성솔루션 채용, (인식범위; VAD, 생활정보, 유머, 영어 등)
	애니넷 서비스	삼성화재	무선 데이터 통신, 휴대폰, 서버기반	-	보이스웨어의 음성솔루션 채용 (인식범위; VAD, 생활정보, 교통상황 안내, POI(목적지; Point of Interest) 검색)
	자동차용 음성 HMI 시스템	산자부 중기거점 프로젝트 연구단	무선 데이터 통신, 휴대폰, 서버기반	-	음성을 이용해서 주변 건물을 찾고 그 결과를 음성을 통해서 받을 수 있는 기능을 시범적으로 구현.

응용 분야	제품명	회사명	채택 기술	표준지원 여부	비고
지능형 로봇 (현재로는 수백 단어 정도의 핵심어 검출이 가능하며 제한된 영역에서의 대화 모델링이 가능한 수준)	URC	ETRI	무선데이터 통신, 서버기반	VoiceXML	URC; Ubiquitous Robot Companion 네트워크를 통한 다양한 정보와 서비스 제공
	가정로봇용 음성명령어 인식엔진	ETRI, 시범서비스	무선데이터 통신, 서버기반	VoiceXML	지능형로봇, 텔레매틱스, 홈네트워크, 차세대PC로 발전가능
	우체국 공공서비스 로봇	ETRI, 시범서비스	무선데이터 통신, 서버기반	VoiceXML	용 우편번호 동명 주소인식 및 대화체 음성합성엔진 탑재
VAD 기능 (Voice Activated Dialing)	V1660 핸드폰	삼성과 스프린트	명령어 음성인식	-	음성으로 휴대폰의 주소록을 검색하여 전화를 걸어주는 서비스 (휴대폰 음성 다이얼링) 보이스시그날사의 내장형 음성인식을 탑재해서 숫자를 연속으로 말해서 전화를 걸도록 하는 기능 내장
모바일 질의응답 시스템(모바일 정보검색)	VAQA	미국	서버기반	-	간단한 대화 기능을 제공하며 TREC 데이터에 대한 질의응답을 수행하는 시스템
	SPIQA	일본	서버기반	-	템플릿 기반의 질문 생성 기능을 갖춘 제한된 영역의 질의응답 시스템(VAQA와 유사)
	AnyQuestion	ETRI,	온톨로지, 텍스트마이닝 기술 서버기반	-	백과사전용의 제한된 질문 패턴 내에서 로봇과 연동하여 인물 및 기록정보에 대한 음성 질의응답을 수행하는 시범 시스템(2005년 말)
통역 시스템	휴대형(PDA) 음성 일영 통역 시스템	NEC	무선데이터 통신, 서버기반	-	2005년 10월
	MASTOR System	IBM	무선데이터 통신, 서버기반 Desktop, PDA	-	2004년, PDA에서 영어, 중국어간 자동 통역 시스템 Tourism, Healthy care & International meeting
	SpeechGuards	ECTACO	"	-	PDA형의 다국어 통역 전용 모바일 단말기 (영어<->불어,독어,이탈리아어, 러시아어, 스페인어, 중국, 일어 등 7개 언어로 자동 통역)(비상용)
	UT103, 203	ECTACO	"	-	영어<->독어, 러시아어 휴대형 통역 단말기(비상용)
	PTW	솔트룩스	"	-	휴대형 음성 통역 시스템, PDA (HCI Lab과 함께)(비상용)
	자동통역 국제간 시연 성공	ETRI(음성언어 정보 연구부)	PSTN, 서버기반	-	1999년 수행(CSTAR 국제공동연구)
번역시스템	PTW	솔트룩스	클라이언트 기반	-	PDA 내장형 기계 번역 시스템 (보이스웨어의 음성엔진을 적용해 번역된 문장 출력, 음성 출력)

응용 분야	제품명	회사	채택 기술 설명	표준지원 여부	비고
플랫폼 개발을 통한 음성인식 정보 서비스	XML 기반 플랫폼, 음성인식 정보 서비스를 ASP 사업화	텔미네트웍	인터넷망, 무선장비, 휴대폰, PDA, 서버기반	VoiceXML	무선장비를 통해 전달되는 서비스의 스케일과 범위를 혁신 시킨 것
	say②me	MPC(콜센터 기반 CRM 전문업체)	인터넷망, PC, 휴대폰, 유무선 단말기, PDA, 서버기반	VoiceXML	VoiceXML기반의 음성인터페이스 서비스 플랫폼을 개발 → KTF와 삼성카드에 VoiceXML 기반 포털 서비스 시스템 공급
	Voice Toolkit (IBM Voice Server)	IBM	PSTN, 인터넷망, 유무선 단말기, 서버기반	VoiceXML, X+V	멀티모덜 지원
	Speech Server	MS	PSTN, 인터넷망, 유무선 단말기, 서버기반	SALT	멀티모덜 지원
	Sand Cherry Soft Server	HP Intel	PSTN, VoIP망, 무선망, 휴대폰, PDA, 텔레매틱스 단말기, 서버기반	SALT, X+V	멀티모덜 지원, 적용 예; Starbucks IVR, 텔레매틱스 서비스에 적용가능
	Oracle AS Wireless	Oracle	PSTN, 유무선 단말기	VoiceXML	
댁내 가전 제어		KT	휴대폰, 유무선 전화	VoiceXML	외부 전화망을 통해 가정 내의 가전제어가 가능한 XML 기반 음성인터페이스 서비스를 개발
멀티모덜 PDA		Loquendo	인터넷망, 클라이언트 방식	X+V, SALT	멀티모덜 지원, PDA내에 TTS, ASR 엔진 장착, Wi-Fi(모바일망)
KIOSK 단말기	MATCHKiosk	AT&T	인터넷망	X+V, SALT	멀티모덜 지원(터치스크린, 펜, 음성)
음성인식/ 합성 엔진	다국어 음성 인식 엔진	Scansoft (구;스피치웍스)	-	-	음성인식 엔진을 개발하여 CTI 분야 및 embedded 분야에서 세계시장 점율 1위
	음성인식 엔진 (PowerASR)	HCI Lab	-	-	휴대폰용 음성 다이얼링 시스템, 얼굴/음성 동시인식 출입관리 시스템, 한-일전화 통역시스템(히타치연구소와 개발중)
	음성인식 엔진 (VSuite2.0)	VoiceSignal	화자독립방식	-	핸드폰 및 모바일 기기에서 사용하는 음성인식 엔진
	음성인식 엔진 (VoiceEZ)	보이스웨어	화자독립/종속방식	-	영어 음성인식/합성기, 중국어 음성 합성기 개발
	음성합성 엔진 (CoreTTS)	코아보이스	-	-	10만 단어 이상의 영어 발음사전 포함, 원격 모니터링
	음성인식 엔진	Philsoft	-	-	
	음성인식 엔진	jamova C.L.S	-	-	대용량 음성인식을 위한 서버용 제품과 4M 미만의 embedded 용 제품
	음성인식 엔진 (VoiceLink)	베스티안파트너스(SL2)	-	-	대용량 음성인식을 위한 서버용 제품과 PC, embedded 기기를 위한 제품으로 구성
	음성인식엔진 (SVOX Speech Server)	SVOX	-	-	embedded 제품군(TTS)과 서버기반 제품군(ASR, TTS)으로 구성
	TTS 엔진	AT&T Natural Voice	-	-	Software Developers Kit
	음성인식 엔진 (Nuance8.5)	Nuance, Elan	-	-	2002, 2003년 업계 1위, 자연어 처리 가능, 28개국 언어 제공

이스를 채택하고 있다. 이들은 단순한 명령어 중심 단어 인식 수준에서부터 표준화(VoiceXML, SALT, X+V)를 지원하는 수준까지 다양하며, 기술적으로 음성인식을 서버에서 혹은 클라이언트에서 할 수 있으며, 음성 정보가 경유하는 망이 전화망 혹은 인터넷망이 될 수 있다는 점에서 다양하게 제품이 구성되어 있다. 아래 표는 현재 개발되어 있는 음성 인터페이스 탑재 예들을 보여주고 있다. 앞으로는 음성을 포함한 멀티 모덜리티를 지원하는 방향으로 제품 개발이 활성화될 전망이다. 아래 예시된 것 이외에도 다양한 분야 서비스(게임, 완구, 가전, 증권 및 날씨 정보 등 제공 서비스 등) 및 회사 제품이 존재한다.

3.1 개발 현황

3.2 음성 인터페이스 구현 유형

지금까지 개발되어 있는 음성 인터페이스 구현 제품의 유형은 대략 5가지로 대별할 수 있다.

- **유형1 (PSTN망을 통해 유무선 전화기를 지원하는 유형)**

이는 유무선 전화 단말기를 통한 음성신호가 PSTN망을 경유함으로써 (웹)서버의 정보를 서비스 받는 방식으로 기존 전화망을 통해 음성정보를 전송하고 VoiceXML(SALT) 게이트웨이에서 음성인식/합성 기능을 처리하는 유형이다. 대표적인 예는 콜센터를 들 수 있다.

- **유형2 (PSTN망 및 인터넷망을 동시에 지원하는 유형)**

이는 PSTN망을 통한 유무선 단말기뿐만 아니라 인터넷 접속가능 단말기를 동시에 지원하는 방식으로 음성처리기능이 서버 혹은 클라이언트에 존재할 수 있다. 그 예로 Sandcherry SoftServer[3]가 있다.

- **유형3 (멀티모덜 DSR 클라이언트-서버 구조)**

이는 멀티모덜 단말기를 사용하는 분산음성인

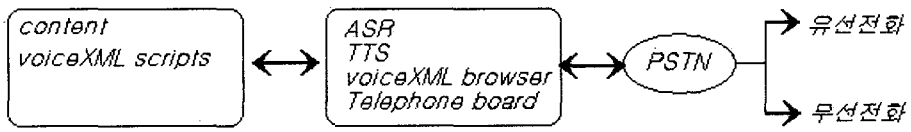


그림 1. 음성 인터페이스 유형1의 개념도

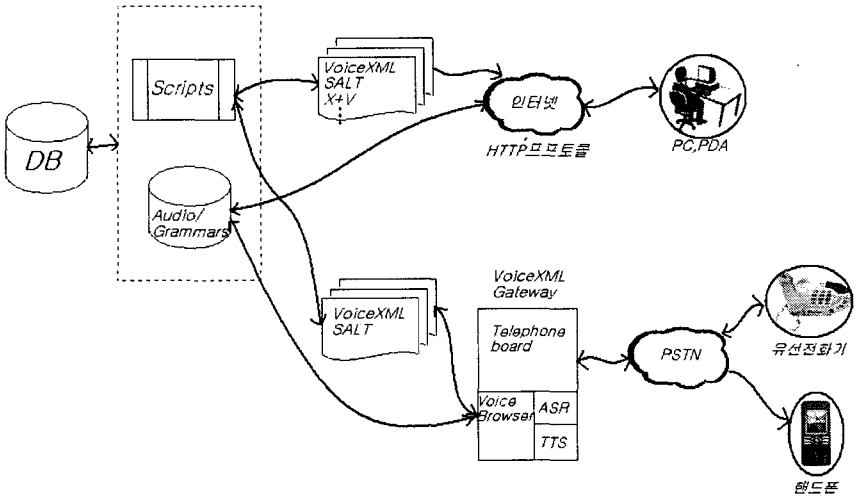


그림 2. 음성 인터페이스 유형2의 개념도

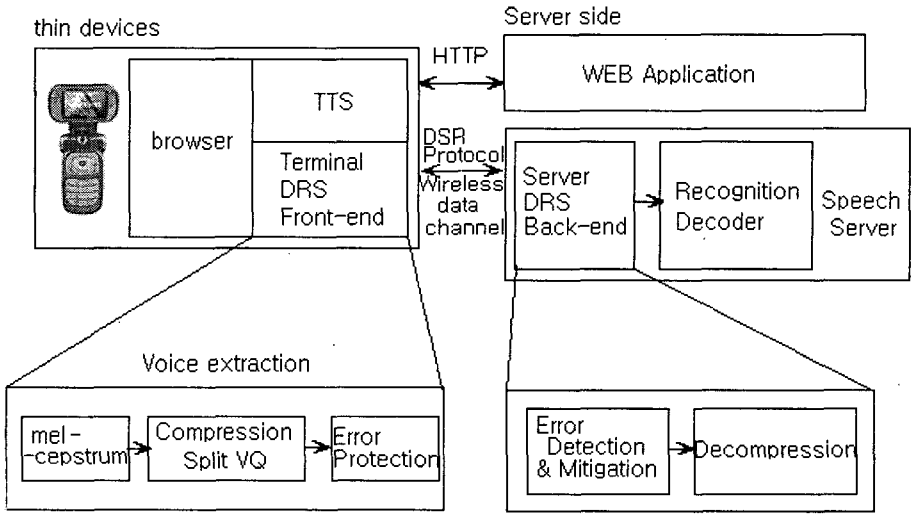


그림 3. 음성인터페이스 유형3의 개념도

식 구조로서 서비스 로직은 서버에 있는 반면에 ASR 처리는 클라이언트와 서버사이에 분산되어 있는 구조이다. 이 방식은 이동통신망 및 단말기 기술이 발전하는 추세에 적합한 구조로서 표준화가 진행 중이다(4장 참고).

• 유형4 멀티모덜 서버 기반 구조

이는 멀티모덜 인터페이스가 매우 단순한 형태에 적합하며 음성인식은 Push-To-Talk(PTT, 무전기 통화방식)에 의해 실행되고 음성의 역할은 단순히 리스트 선택을 위해 사용되는 반면에 확인은 단말기의 버튼들을 눌러서 하는 방식이다. 이는 완전히 서버 기반 모델이라 할 수 있다. 현재로서 가장 보편적인 멀티모덜 지원 방식이다.

• 유형5 멀티모덜 클라이언트 기반 구조

PDA나 태블릿 PC, Kiosk, 차량 단말기, 노트

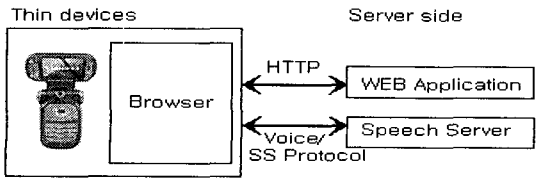


그림 4. 음성인터페이스 유형4의 개념도

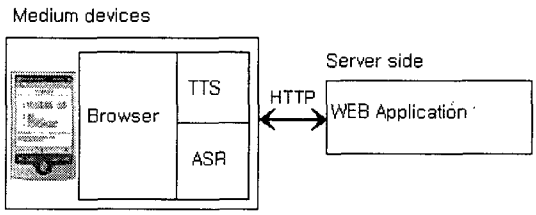


그림 5. 음성인터페이스 유형5의 개념도

북, 데스트탑 같은 중간크기 단말기에서는 메모리 및 처리 능력 차원에서 클라이언트 기반의 구조가 채택될 수 있다. 여기서는 임베디드 TTS 및 ASR 엔진이 단말기에 장착되며, 서버에 대한 요구 수나 응답으로서의 스트리밍 커넥션이 줄어드는 방식이다. 이동통신이 발전할수록 멀티모덜 단말기의 성능이 좋아지게 되며 이때 적합한 방식이라 할 수 있다.

4. 음성인터페이스 관련 국제표준화 동향

음성인터페이스 관련 표준화는 음성 대화, 음성인식/합성, 정보시스템, 전화망 등의 접속망을 상호 분리하여, 음성 정보시스템 구성요소들 각각의 상호 독립적인 개발을 보장해 주며, 각 요소의

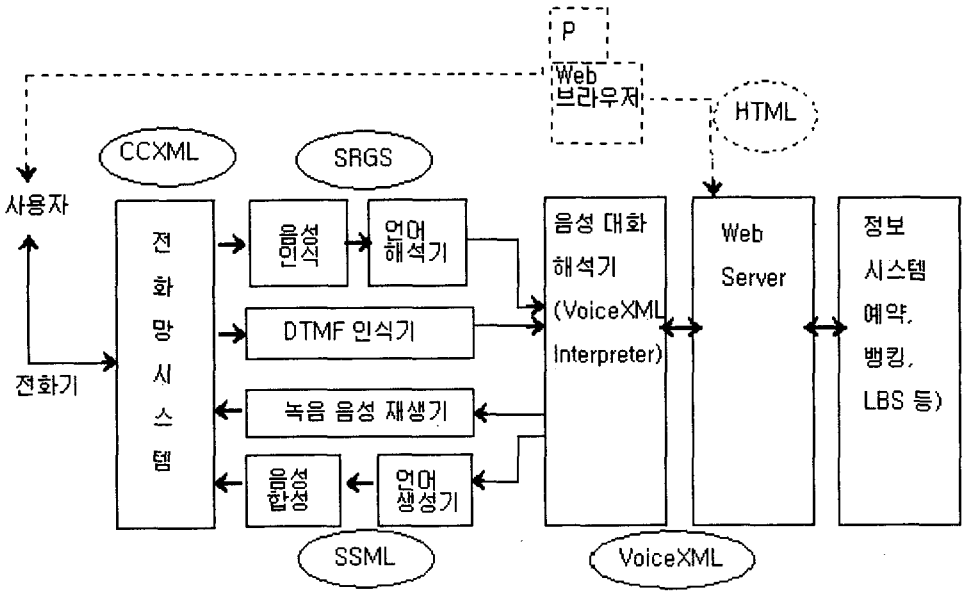


그림 6. W3C의 음성 브라우저 개념도

이해가 없이도 음성 대화만을 설계 기술하여 음성정보시스템을 개발 할 수 있도록 함으로써 음성정보기술의 보급 및 확산에 크게 기여하고 있다[4]. 표준화동향은 다음과 같다.

4.1 함수 수준(API)의 표준화

음성엔진 API로는 JSAPI(Java Speech API), JTAPI(Java Telephony API), MS SAPI(Speech API) 등이 개발되었다. 이 표준으로 인해 이 API의 숙지만으로 음성대화 시나리오를 개발할 수 있다.

W3C에서는 음성 및 멀티모덜 영역의 표준화를 리드하고 있다. Voice Browser WG은 1998년에 생겨서 웹기반의 음성 관련 표준화(음성 응용[VoiceXML], 음성 인식[SRGS]을 진행하고 있으며, MultiModal Interaction WG(http://www.w3.org/2002/mmi)은 2002년에 생겨서 웹 관점에서 통합된 멀티모덜 응용을 개발하기 위해 프레임워크[5]를 만들고 있다.

4.2 W3C의 웹 기반 음성 브라우저(Voice Browser) 표준화 활동

위 그림은 W3C 음성 브라우저의 구조를 보여주고 있다. **SSML**(Speech Synthesis ML; 2004년 9월 8일 표준 확정)[6]은 음성 합성 문법 명세서로서 TTS를 단순하고 표준적으로 제어하게 할 수 있으며, **SRGS**(Speech Recognition Grammar Specification; 2004년 3월 16일 표준 확정)[7]은 음성 인식 문법 명세서이며 음성과 DTMF(touch tone) 입력 문법을 모두 기술할 수 있으며 이는 ASR 제품 벤더들의 의해 채택되었으며 음성인식문법을 코딩하기 위한 XML 포맷과 텍스트 포맷(ABNF 포맷)을 포함하고 있으며, **CCXML**(Call Control XML; 2005년 6월 29일 Last Call WD)[8]은 전화망 제어와 관련한 명세서이며, **VoiceXML**(VoiceXML 2.0; 2004년 3월 16일 표준 확정)[9]은 전화통신을 기반으로 하는 상호 주도형 음성 대화(Dialogue)를 기술하는 명세서이다. 이러한 XML 기반의 음성대화 및 인식/합성 마크업은 웹 기반 컨텐츠의 음성

```
<?xml version="1.0">
<!DOCTYPE vxml SYSTEM "vxml.dtd">
<vxml version="2.0">
  <form>
    <var name="h" expr="1"/>
    <var name="m" expr="34"/>
    <var name="s" expr="19"/>
    <field name="r" type="boolean">
      <prompt>VXML만으로는 현재시각을 알 수 없습니다. 미리 입력된 시간은
        <vale expr="h"/> 시,
        <vale expr="m"/> 분,
        <vale expr="s"/> 초입니다.
      <prompt> 다시 들으시겠습니까?
      </prompt>
      <filled>
        <if cond="hear_another">
          <clear/>
        </if>
      </filled>
    </field>
  </form>
<vxml>
```

그림 7. VoiceXML 샘플

서비스 응용을 개발 가능하게 해준다. VoiceXML 2.0은 W3C에서 IBM 주도로 Motorola, AT&T, Lucent Technologies 등이 전화망 음성인터페이스 응용을 위해 설계된 표준안을 발표하였다. XML 기반의 음성인터페이스는 현재 단어와 단문을 인식할 수 있는 단계이며, 향후 자연스러운 대화체 인식 및 합성이 가능해질 것이다.

VoiceXML 3.0은 진보된 음성정보시스템에서 사용될 수 있는 효과적인 대화 기술을 가능하게 하고, 다른 W3C 언어와 쉽고 명확하게 통합할 수 있는 형태를 제공하는 것이 목표이다. 멀티모덜 인터페이스를 위해 다양한 마크업 언어(XHTML, SMIL, SALT, WAI 등)간의 통합이 가능해야 할 것이다.

4.3 음성을 기반으로 한 멀티모덜 인터페이스의 표준화 작업

- MS 주축의 SALT
- IBM 주축의 XHTML+Voice
- W3C의 멀티모덜 상호작용(MultiModal Interaction)

멀티모덜 영역에서는 2개의 산업체로 구성된 포럼이 활동 중이다. VoiceXML 포럼(http://www.voicexml.org)과 SALT 포럼(http://www.saltforum.org)이며, 전자는 VoiceXML을 제안한 이후 현재는 XHTML+VoiceXML(일명 X+V)이란 멀티모덜 규격 언어를 지원하고 있으며, IBM, Motorola, Opera 등이 참여하고 있다. 후자는 Microsoft, Comverse, Cisco, Philips, Scan Soft, Intel 등에 의해 발족되어 SALT를 개발하였으며, 음성을 포함하는 멀티모덜 인터페이스 표준화를 추진하고 있다. 이는 HTML/ XHTML, WML, SMIL 페이지에 음성을 통합하기 위해 추가된 작은 태그 집합이다.

- **SALT(Speech Application Language Tags)**

SALT는 다른 마크업 언어(HTML, XHTML, WML)의 확장으로 음성만을 위한 브라우저와 멀

티모덜 브라우저의 2가지 목표를 달성하기 위한 명세이다. 기본적으로 PC용 웹브라이저를 위한 HTML(또는 XHTML)이나, 휴대전화 또는 PDA용 웹브라우저를 위한 WML에 내포될 수 있도록 설계되었으며, 음성명령으로 소프트웨어를 제어할 수 있게 하였다. 따라서 비주얼 페이지를 통해 음성 입출력을 동시에 지원할 수 있는 멀티모덜 인터페이스를 기술할 수 있는 언어 명세이다. 음성인식 문법은 W3C의 SRGS를 그대로 사용한다. MS사는 Visual Studio.Net과 인터넷 익스플로러 .Net initiative에 SALT 호환 음성 인터페이스 엔진을 개발, 제공할 계획이다.

• **X+V**

IBM이 주도하는 컨소시엄에 의해 개발된 X+V[10]는 SALT[11]와 아주 유사하며, 단지 X+V에서는 음성대화관련 마크업을 W3C의 VoiceXML을 그대로 사용하는 반면, SALT에서는 별도로 개발하여 사용하고 있다는 차이점이 있다. X+V에서는 XHTML을 호스트언어로 하고, VoiceXML을 내포 언어로 채용하고 있다.

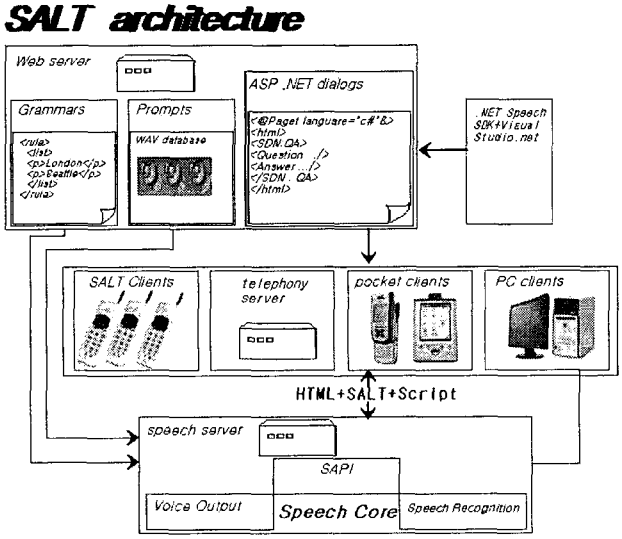


그림 8. SALT의 개념도

```
<!-- HTML -->
<html xmlns:salt="http://www.saltforum.org/2002/SALT">
  <form id="travelForm">
    <input name="txtBoxCity" type="text" />
    <input name="buttonCityListen" type="button" onClick="listenCity.Start();" />
  </form>

<!-- SALT -->
  <salt:listen id="listenCity">
    <salt:grammar name="g_city" src="./city.grxml" />
    <salt:bind targetelement="txtBoxCity" value="//city[1]" />
  </salt:listen>
</html>
```

그림 9. SALT 샘플

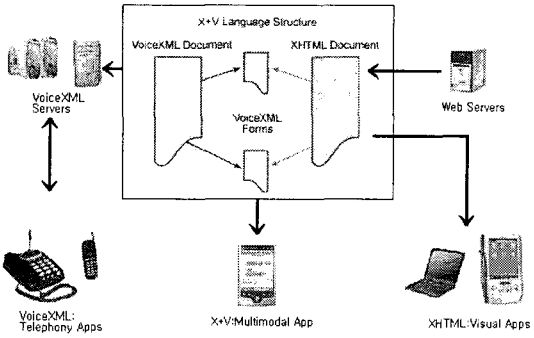


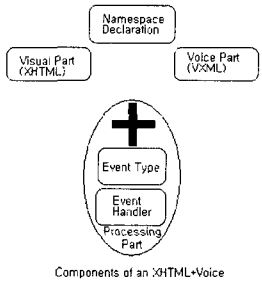
그림 10. X+V의 개념도

X+V와 SALT는 W3C의 멀티모덜 인터페이스 표준으로 채택되기 위해 제출된 상태이며 이 둘은 서로 경쟁관계에 있다.

- **W3C의 멀티모덜 상호작용(MultiModal Interaction)**

최근 음성처리 기술, 필기체 인식 기술이 발전하고, 성능이 개선된 지능형 초소형 단말기가 출현하면서 다양한 입력을 처리하는 새로운 인터페이스에 대한 연구가 본격적으로 진행되고 있으며 대표적인 연구가 멀티모덜 인터페이스이다. 이는 음성, 시각, 촉각 등과 같은 단일 모덜리티를 복합적으로 이용하여 보다 편리하고 쉽고 정확하게 컴퓨터와 대화하는 것을 의미한다. W3C는 멀티모덜 기술을 활용하는 서비스가 요구됨에 따라, 멀티모덜 인터랙션 워킹그룹(MultiModal Interaction WG)을 만들어 다양한 인터페이스 모드를 지원하기 위한 표준의 개발을 목표로 표준화 활동을 진행하고 있다. 멀티모덜 인터페이스는 음성 입출력을 처리하는 음성 인터페이스, 펜 필기체를 인식하는 잉크 인터페이스 및 보편적인 GUI를 모두 포함하고 있으며, 앞으로 새로이 발전하는 인터페이스 기술들을 추가 가능하다.

현재 MultiModal Interaction Framework (2003년 1월 8일 요구사항 분석 상태)은 적용시스템의 일반적인 구조와 구성요소 및 사용할 수 있는




```
<?xml version="1.0"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD
  XHTML+Voice //EN" "xhtml+voice10.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
  xmlns:vxml="http://www.w3.org/2001/voicexml20"
  xmlns:ev="http://www.w3.org/2001/xml-events">
  <head>
    <title>What You See Is What You Can Say</title>
    <!-- first declare the voice handlers. -->
    <vxml:form id="voice_city">
      <vxml:field name="field_city">
        <vxml:grammar src="city.srgf"
          type="application/x-srgf"/>
        <vxml:prompt id="city_prompt">
          Please choose a city.
        </vxml:prompt>
        <vxml:catch event="help nomatch noinput">
          For example, say Chicago.
        </vxml:catch>
      </vxml:field>
    </vxml:form>
    <vxml:form id="voice_hotel">
      <vxml:field name="field_hotel">
        <vxml:grammar src="hotel.srgf"
          type="application/x-srgf"/>
        <vxml:prompt id="hotel_prompt">
          Select your hotel
        </vxml:prompt>
        <vxml:catch event="help nomatch noinput">
          For example, say Hilton.
        </vxml:catch>
        <vxml:filled>
          <vxml:prompt>
            You have chosen to stay at the
            <vxml:value expr="field_hotel"/>
          </vxml:prompt>
        </vxml:filled>
      </vxml:field>
    </vxml:form>
  <!-- done voice handlers. -->
  </head>
  <body>
    <h1>What You See Is What You Can Say</h1>
    ..............
</html>
```

그림 11. X+V 샘플

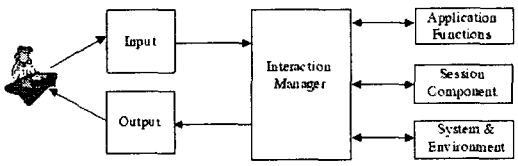


그림 12-(a). 멀티모달 인터랙션의 개념도

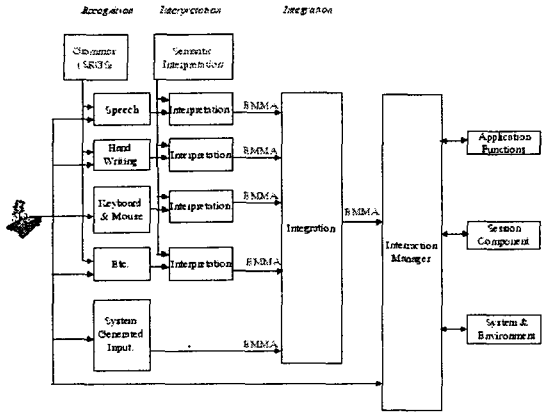


그림 12-(b). 멀티모달 인터랙션의 개념도(입력부)

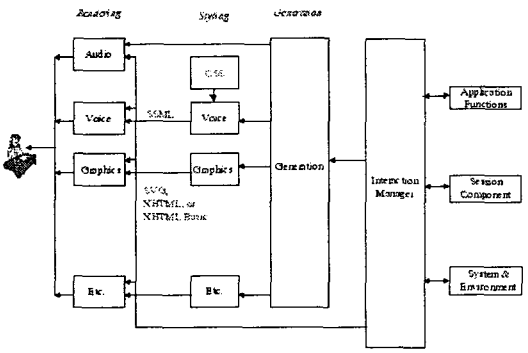


그림 12-(c). 멀티모달 인터랙션의 개념도(출력부)

표준 및 마크업 언어와 관계를 명기하고 있으며, EMMA(Extensible MultiModal Annotation ML; 2005년 9월 16일 Last Call WD 상태)[12]는 입력 장치들과 멀티모달 상호작용 관리 시스템 사이의 인터페이스를 위한 데이터 형식을 표준화하기 위한 것이며, 펜입력(InkML; 2005년 4월 Last Call WD 상태)(http://www.w3c.org/TR/inkML/)은 멀티모달 시스템의 전자펜이나 스타일러스에서 사용되는 잉크를 위한(필기체를 인식한 결과를 표현해 주는 언어) XML 형식을 정의하고 있다.

MultiModal Interaction Framework의 기본요소는 다음과 같다. "입력요소"는 그림과 같이 다양한 사용자의 입력을 받아서 인식모듈, 해석모듈(인식 결과의 의미[semantic] 정보를 판단)을 거쳐서 EMMA(사용자의 명령 및 모덜리티 정보를 포함하는 데이터)를 생성한다. 다양한 정보(예; 음성과 필기체, 제스쳐, 디지털 INK 포맷 등)는 통합모듈(하나의 데이터로 통합)을 거쳐 인터랙션 매니저로 전달된다. "출력요소"는 생성모듈(인터랙션 매니저가 출력 정보를 어떤 모덜리티로 출력할지를 결정), 스타일 모듈(각 모덜리티에 대한 스타일 정보 기술) 및 렌더링 모듈(스타일 정보에 따라 다양한 포맷으로 출력)로 나누어진다. "인터랙션 매니저"는 입력요소로부터 얻은 EMMA 데이터를 이용하여 실제 응용 서비스를 실행한 후, 그 결과를 출력 요소에 제공하며, 다양한 모덜리티의 인터랙션을 통합하며, 모덜리티간의 동기화를 이룬다. "세션요소"에서는 세션이 끊어지지 않도록 지속적인 연결 및 다양한 단말기 출력을 위한 싱크 기능을 한다. "시스템 환경요소"는 디바이스 선정 및 사용자 선호에 따라 자동으로 변화하도록 시스템 환경을 표현해준다. "응용 기능"은 응용 서비스 실행 요소이다.

EMMA는 입력 요소의 결과를 표현하는 표준화된 교환 포맷으로서 3가지 구성요소(Instance Data[의미데이터], Data Model[데이터모델], Metadata[메타데이터])로 이루어지며, 의미데이터는 해석모듈에 의해 해석된 결과로써 XML 문서로 기술되며, 데이터모델은 의미 데이터의 DTD이며, 메타데이터는 속성(입력 모덜리티 종류, 입력 시간, 입력 종료 시간, 인식결과의 신뢰값 등)을 의미한다.

진행되고 있는 멀티모덜 인터페이스 관련 서비스로는 버튼 스위치, 터치 스크린, 음성을 입력으로 사용하여 특정 목적지를 찾아가는 텔레매틱스 서비스를 생각할 수 있으며, 이때 출력은 오디오, 음성 및 그래픽 형태가 될 수 있다. 또 다른 서비스로 휴대폰을 통한 멀티모덜 다이얼링 서비스 혹은 기차표 예약 등을 예로 들 수 있다. 이때 전화 및 예약을 한때 음성, 키패드, 펜 등을 입력수단으로 할 수 있다.

4.4 분산처리를 위한 표준화

유비쿼터스 환경과 같이 분산 환경이면서, 다수의 이기종 클라이언트, 서버를 활용해야 하는 경우의 음성 정보 시스템을 위하여 다음과 같은 표준이 진행 중이다.

• MRCP

MRCP[13]는 IETF(Internet Engineering Task Force; http://www.ietf.org/)의 Speech Control WG에 의해 진행되고 있는 분산 음성 처리 및 응용을 위한 통신 프로토콜이다.

이는 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. 클라이언트와 서버가 주고받는 메시지는 SRGS, SSML 등을 이용하도록 설계되어 있다.

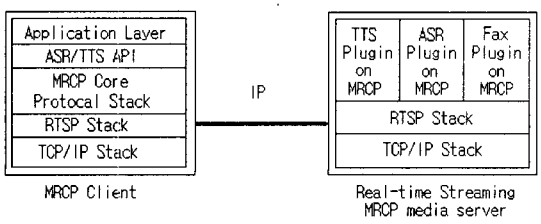


그림 13. MRCP의 개념도

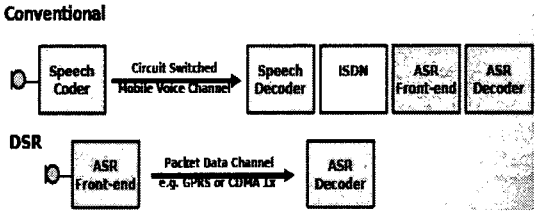


그림 14. 기존 방식 DSR의 개념도 비교

• ETSI standard DSR front-end

유럽 ETSI DSR Aurora WG[14]에서 2000년부터 음성인터페이스 전처리부를 표준화하였으며, 휴대폰 통신단말기에서 음성전처리만 수행하고 음성인식은 서버를 사용하는 DSR 방식의 음성인터페이스를 개발하였다.

모바일 환경 하에서 PDA 단말기의 자원을 최소로 사용하기 위하여 패킷 기반의 음성 정보처리를 수행하는 분산음성인식(DSR) 기술이 사용되었다. 이 기술은 무선 랜에서의 패킷 손실에 의한 음성인식 성능 저하를 막기 위해 단말기에서는 잡음제거와 음성의 특징만을 추출하여 전송을 하고, 서버에서 이 특징을 기반으로 음성인식 처리를 하는 방식으로 스마트폰 및 모바일 폰 등의 무선 환경에서의 음성처리에 필수적인 기술이다. 임베디드 TTS 기술은 단말기에 존재하고, 반면에 ASR 처리는 클라이언트와 서버에 분산된다. 반면에 훨씬 단순한 단말기의 경우 그림4와 그림14의 상단 그림과 같이 서버 기반 구조로 구성될 수 있다. ETSI(http://portal.etsi.org) DSR 표준에는 그림3과 같이 분산음성인식을 위한 MFCC 계산, 양자화, 오류 처리 등에 대한 틀이 제공된다.

5. 결론 및 향후 방향

본문에서 언급한 개발현황을 볼 때 멀티모덜 관련 표준화를 기반으로 하는 음성 정보시스템의 개발은 국내에서는 아직 미진한 상태인 반면에

국외에서는 VoiceXML 및 SALT를 기반으로 하는 음성정보 응용이 매우 빠르게 확산되고 있다. 국내에서도 유비쿼터스 환경에 적합한 멀티모덜 인터페이스의 표준에 따르는 음성 인터페이스의 개발이 요구되고 있는 실정이다.

참 고 문 헌

[1] 이윤근, 박준, 김상훈, "정보통신 미래기술 특집 (1.음성 인터페이스 기술)", 전자통신동향분석 제20권 제5호(통권 95호), pp1~16, ETRI(한국전자통신연구원), 2005년 10월.

[2] Paolo Baggia, Silvia Mosso, "Speech Technologies and MultiModality: The Solution for New Advanced Services", Loquendo White Paper, April 28, 2005.

[3] http://www.sandcherry.com/products, Sandcherry Inc., "Extending Applications to Voice and Multimodal Users", Part Number:BP28-2003/E SB0361, 2003.

[4] 김병수 외 다수, "특집: 유비쿼터스 환경의 음성언어기술", 한국정보과학회지, 제24권 제1호 통권 제200호, pp7~67, 한국정보과학회, 2006년 1월.

[5] W3C MultiModal Interaction Framework, http://www.w3.org/TR/mmi-framework/, W3C Note, May 2003.

[6] D.C.Burnett et al., "SSML Version 1.0", W3C Recommendation, 7 Sep. 2004, See: http://www.w3c.org/TR/Speech-synthesis/.

[7] Andrew Hunt, Scott McGlashan, "Speech Recognition Grammar Specification Version 1.0", W3C Recommendation, 16 March 2004. See: http://www.w3c.org/TR/ speech-grammar/.

[8] R. J. Auburn, "Voice Browser Call Control: CCXML Version 1.0", W3C Last Call Working Draft, 11 Jan. 2004. See: http://www.w3c.org/TR/ccxml/.

[9] Scott McGlashan et al., "Voice Extensible Markup Language(VoiceXML) Version 2.0", W3C Recommendation, 16 March 2004. See: http://www.w3c.org/TR/ voicexml20/.

[10] Chriss Cross et al., "XHTML+Voice Profile 1.1", 28 Jan. 2003. See: http://www- 306.ibm.com/software/pervasive/multimodal/x+v/11/spec.htm.

[11] SALT Forum, "Speech Application Language Tags(SALT 1.0) Specification", 15 July 2002. See: http://www.saltforum.org/saltforum/downloads/SALT1.0.pdf.

[12] Wu Chow et al., "Extensible MultiModal Annotation markup language", W3 Working Draft, 14 Dec. 2004. See: http://www.w3c.org/TR/emma/.

[13] Saravanan Shanmugham, et al., "Media Resource Control Protocol Version2 (MRCPv2)", Internet Draft of IETF Speech Control Working Group, May 2004.

[14] ETSI ES 201 108 version 1.1.2, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; front-end feature extraction algorithm; Compression algorithm", April 2000.

장 민 석

1989년 연세대학교 전자공학과 공학사
1991년 연세대학교 전자공학과 공학석사
1997년 연세대학교 전자공학과 공학박사
1997년~현재 군산대학교 컴퓨터정보과학과 부교수
관심분야 : 웹기반기술, 프로토콜공학, 소프트웨어공학

김 성 국

2004년 군산대학교 컴퓨터정보과학과 이학사
2004년~현재 군산대학교 컴퓨터정보과학과 이학석사 과정중
관심분야 : 웹기반기술(VoiceXML, SALT)