VoiceXML 기반 음성인식시스템을 이용한 서비스 개발

김학균(KT), 김은향(KT), 김재인(KT), 구명완(KT)

<차 례>

<Abstract>

The Interactive Voice Services based on VoiceXML

Hak-Gyoon Kim, Eun-Hyang Kim, Jae-In Kim, Myoung-Wan Koo

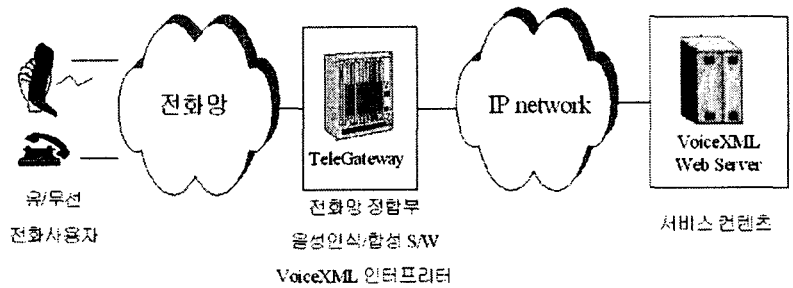
As there are needs to search the Web information via wire or wireless telephones, VoiceXML forum was established to develop and promote the Voice eXtensible Markup Language (VoiceXML). VoiceXML simplifies the creation of personalized interactive voice response services on the Web, and allows voice and phone access to information on Web sites, call center databases. Also, it can utilize the Web-based technologies, such as CGI(Common Gateway Interface) scripts. In this paper, we have developed the voice portal service platform based on VoiceXML called TeleGateway. It enables integration of voice services with data services using the Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) engines. Also, we have showed the various services on voice portal services.

* 주제어: VoiceXML, TeleGateway, VAD

1. 서 론

음성 인식 및 합성기술을 기반으로 한 대화형 음성언어 인터페이스는 시각에 의존하는 기존의 웹을 벗어나 음성 및 시각을 모두 활용할 수 있는 새로운 패러다임을 제시하고 있다[1]. 그러나, 기존의 음성서비스는 인터넷의 다양한 정보를 제대로 활용하지 못하며, 시나리오를 작성하는 표준 언어가 없기 때문에 동일한 시나리오도 서비스 플랫폼 제작 회사에 따라 다른 s/w로 새로 만들어야 하는 불편한 점이 있었다. 이에 음성서비스 시나리오의 표준으로 제시되고 있는 VoiceXML의 등장으로 다양한 정보 제공과 웹 기반 기술을 활용할 수 있게 되었다.

정보 제공자는 VoiceXML 표준에 따라 음성 입출력이 가능한 시나리오를 만들고, 장비 제공자는 서비스 시나리오에 독립되게 장비를 구축할 수 있으므로 다양한 음성서비스가 신속하게 개발될 수 있다. 유무선 전화사용자는 TeleGateway의 VoiceXML 인터프리터에 의해 구동되는 VoiceXML 문서의 시나리오에 따라 음성서비스를 제공받을 수 있다. 그림 1은 대화형 음성언어 인터페이스를 개략적으로 구성한 것이다.



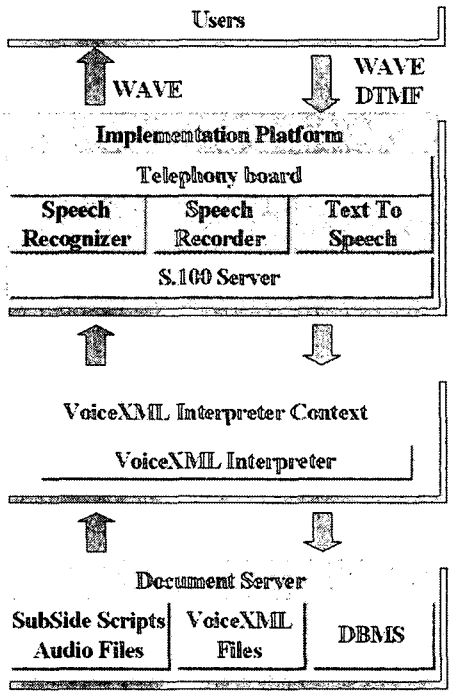
<그림 1> 대화형 음성언어 인터페이스 구성도

본 논문은 유무선 전화 사용자와 정보제공자를 연결해주는 TeleGateway에 대해 설명하고, VoiceXML에 포함되어 있는 다이알로그 모델, 문법 표현 등의 요구사항을 정리한다. 또한, TeleGateway를 활용한 VAD(Voice Activated Dialing) 및 개인전화번호 관리 서비스를 제시한다.

2. TeleGateway

TeleGateway에는 VoiceXML 인터프리터와 전화망 인터페이스 카드 및 음성인식, 합성 소프트웨어가 설치되어 있다. 이것은 일반 유무선 전화사용자의 호에 따라

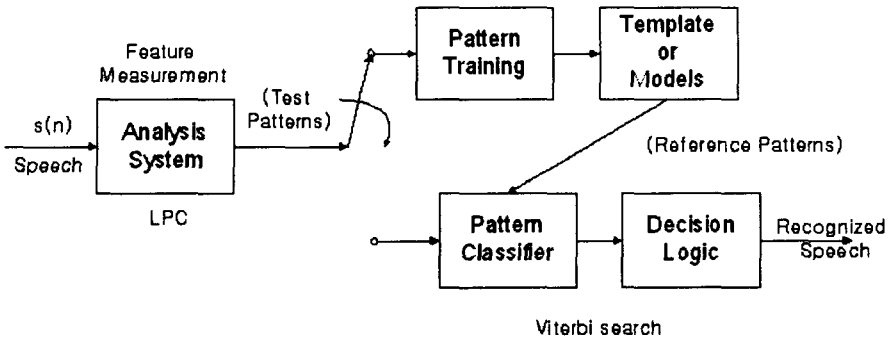
해당 문서의 시나리오를 구동하여, 음성 입출력이 가능한 대화형 음성언어 인터페이스 환경을 제공한다. 구조는 그림 2와 같다.



<그림 2> TeleGateway 구조도

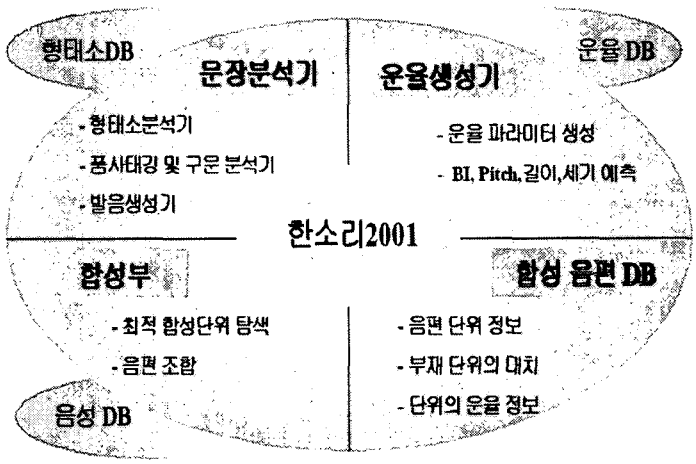
2.1. 음성 인식기

음성 인식기는 수집된 훈련용 음성 데이터를 이용하여 생성한 템플릿 또는 모델을 가지고 있으며, 사용자가 발화한 음성을 등록된 어휘들과 비교하여 가장 비슷한 어휘를 결과로 내놓는다. 따라서 음성 인식기를 구현하기 위해서는 언어 모델들을 생성하기 위한 훈련 과정이 필요하지만, 클라이언트 프로그램에서는 단지 그 결과만 사용하므로 훈련 과정과는 직접적인 관련이 없다. Analysis System은 음성의 특징을 추출하는 기능을 가지고 있으며, LPC(linear predictive coding) 계수들이 사용되었다. 그림 3은 음성인식기의 간략한 구조를 보여준다. 인식된 결과를 확인하기 위해서 어휘검증(utterance verification) 단계를 거친다[3].



<그림 3> 음성 인식기 구조도

2.2. 음성 합성기



<그림 4> 무제한 음성 합성기의 구성도

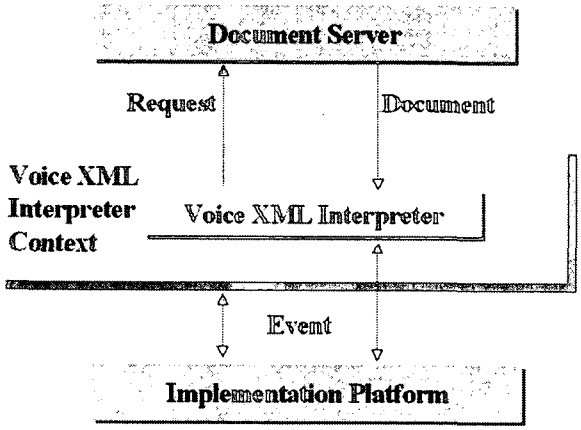
한소리(Hansori)[5]는 KT가 개발한 코퍼스 기반 무제한 음성 합성기로써, 한국어를 음성으로 바꿔주는 역할을 한다. 코퍼스 기반 음성합성기는 사람의 목소리의 변형 없이 음색을 살려 자연스러운 음성으로 합성한다는 장점을 가지고 있으며, 반면에 가능한 모든 운율을 표현할 수 없기 때문에 운율이 부자연스러운 문장이 합성될 수 있다는 단점을 안고 있다. 합성기 한소리는 녹음할 텍스트 선정에서부터 운율을 고려하였고, 녹음 후 DB의 음성이 각 운율요소들의 특징을 고루 포함하도록 하는 작업과, 부자연스러운 운율을 방지하기 위한 기법 등을 적용하여 이러한 단점을 최대한 보완하였다. 다음 그림 4는 음성 합성기의 구조를 보여준다.

2.3. Telephony Interface

TeleGateway는 전화망 인터페이스 카드로서, 보통 안내음성 에코 제거가 가능한 상용제품을 사용하는데, 그 이유는 사용자가 안내 방송이 출력되더라도 아무 때나 음성입력을 할 수 있어 서비스 사용상 편리하기 때문이다.

2.4. VoiceXML 해석기

다이알로그 매니저로서 VoiceXML 해석기는 어플리케이션의 플로우를 제어하며, 특정 시간에 어떤 모듈(ASR, TTS, 데이터 검색 등)을 사용해야 할 지를 결정한다. 또한, 변수값, 실행 정보(execution stack)등과 같은 상태 정보를 이용하여 서로 다른 타스크로의 변동이 가능한 구조를 제공한다.



<그림 5> VoiceXML 아키텍쳐 모델

3. VoiceXML

VoiceXML은 AT&T, Motorola, Lucent Technology 등으로 구성된 VoiceXML Forum이 제시한 음성 입출력 기반의 음성서비스 시나리오의 표준이며, 음성 인식 및 합성기술을 이용하여 인터넷의 풍부한 정보를 누구나 쉽게 음성 입출력이 가능한 대화형 음성서비스 형태로 제공이 가능해진다. 또한, 웹 기반 기술을 음성서비스에 접목시킬 수 있으며, 다양한 서비스를 쉽고 빠르게 제공할 수 있는 환경을 제공한다.

그림 5는 VoiceXML의 아키텍쳐 모델을 보여준다. Document server 즉, 웹 서버

는 VoiceXML 인터프리터의 요청을 받아 VoiceXML 문서를 보내주며, VoiceXML Interpreter Context는 사용자의 입력 및 플랫폼에 종속되는 이벤트를 받는 역할을 한다. 이와 같이 구성된 아키텍처에 기반하여 인터프리터는 문서를 해석하여 적절한 행동을 취하는 구조를 가진다[2].

기존의 음성서비스 시나리오가 제공하는 기능과 다양한 형태의 문서 구조, 다이알로그 모델 등을 지원하기 위해서 VoiceXML은 다양한 엘리먼트를 제공한다. 표 1은 VoiceXML에서 제공하고 있는 엘리먼트를 분류하여 정리한다.

<표 1> VoiceXML의 엘리먼트

분류 항목	상세 내용
Event Handling	<noinput>, <nomatch>, <help> 등의 이벤트들을 처리하고 throw 해주는 엘리먼트 집합
Executable Contents	변수 선언, 제어문 등을 처리하는 엘리먼트 집합
Forms	Form item들을 논리적 단위로 묶는 다이알로그 엘리먼트
Form Items	사용자의 입력 등을 받아들이는 다이알로그 요소 집합
Grammars	음성 인식기에서 사용되는 그래마 지정 및 DTMF를 표현하는 엘리먼트 집합
Links	<link>의 범위에 있는 모든 다이알로그에 적용되는 전이(transition) 룰을 표현
Menus	다이알로그의 한 형식인 메뉴를 표현하는 엘리먼트 집합
Prompts	음성 합성기 혹은 출력단에 보내질 프롬프트를 가공, 출력하는 엘리먼트 집합

3.1. 문서 구조 (Document Structure)

VoiceXML 어플리케이션은 다음과 같은 세 가지로 구분된다.

- 단일 문서(Single document application)
- 멀티 문서(Multi document application)
- 서브 다이알로그(Subdialog)

단일 문서 어플리케이션은 단일 문서로 서비스 시나리오를 구성한 것이다. 멀티

문서 어플리케이션은 루트 문서에 연결된 여러 개의 하위 문서들이 서로 정보를 공유하여 하나의 시나리오를 구성하며, 하위 문서들은 루트 문서의 다이알로그를 공통적으로 사용할 수 있다. 서브 다이알로그는 일종의 서브루틴 개념으로써, 일반적으로 자주 사용되는 다이알로그를 독립된 문서로 만들어 여러 문서에서 다이알로그를 재 사용할 수 있는 환경을 제공한다.

3.2. 다이알로그 모델

VoiceXML는 computer directed form과 mixed initiative form 을 지원한다. form 은 하나의 논리적인 대화 묶음인 다이알로그 형식으로써, <form>, <menu> 두 종류가 있다.

3.2.1. Computer Directed Forms

간단하면서 가장 흔하게 사용되는 다이알로그 모델로서, 시나리오는 미리 정의된 순서에 따라 컴퓨터와 사용자가 대화하는 형식을 가지기 때문에, 다이알로그 요소(form item)는 일정한 순서에 따라 실행된다.

3.2.2. Mixed Initiative Forms

computer directed와는 달리 컴퓨터와 사용자 모두 대화의 진행을 능동적으로 변경할 수 있는 다이알로그 모델로서, form 레벨의 그래마를 필요로 한다. 이것은 사용자의 응답에 따라서 다이알로그 요소가 실행되는 순서를 변경시키거나 여러 개의 다이알로그 요소에 적용시킬 수 있는 기능을 부여한다.

3.3. Form Interpretation Algorithm

VoiceXML은 Form Interpretation Algorithm(FIA)을 이용하여 DOM(Document Object Model) 구조를 순회하면서 서비스 시나리오를 해석한다[6]. 즉, FIA는 각각의 다이알로그(form)의 해석방법을 정의하는 것이다. FIA의 알고리즘은 다음과 같다.

3.3.1. Initialization

각각의 form을 방문할 때마다, form level의 일반 변수 및 form item 변수를 초기화한다. 다음부터 소개될 selection, collection, processing 단계는 모든 form item이 방문 될 때까지 수행한다.

3.3.2. Selection

다음에 처리할 form item을 선택하는 단계로서, 그 역할은 다음과 같다. 다음에 처리될 것을 지정되었을 경우에는 지정된 form item을 선택하고, 아닐 경우 처리되지 않은 form item을 선택한다. 모든 form item이 처리되었을 때는 루프를 종료한다.

3.3.3. Collection

사용자 프롬프트를 내보고, 현재의 form item에서 적용되는 그래마를 활성화시킨다. 그리고 전화 사용자의 입력이나 이벤트를 기다린다.

3.3.4. Processing

collection 단계에서 받은 사용자의 입력이나 이벤트를 처리하는 단계로서 다음과 같은 작업을 한다.

- 전화가 끊기는 등의 특정 이벤트가 발생할 경우, 해당 이벤트를 핸들링 한다.
- 사용자의 입력이 현재 form 이외의 form에 적용될 경우 해당 form으로 제어권을 넘긴다.
- 사용자의 입력이 현재 form에 적용될 경우, 입력 값을 해당 form item에 할당하고, 해당 item의 <filled> 액션을 수행한다.

3.4. 문법 표현(Grammar Representation)

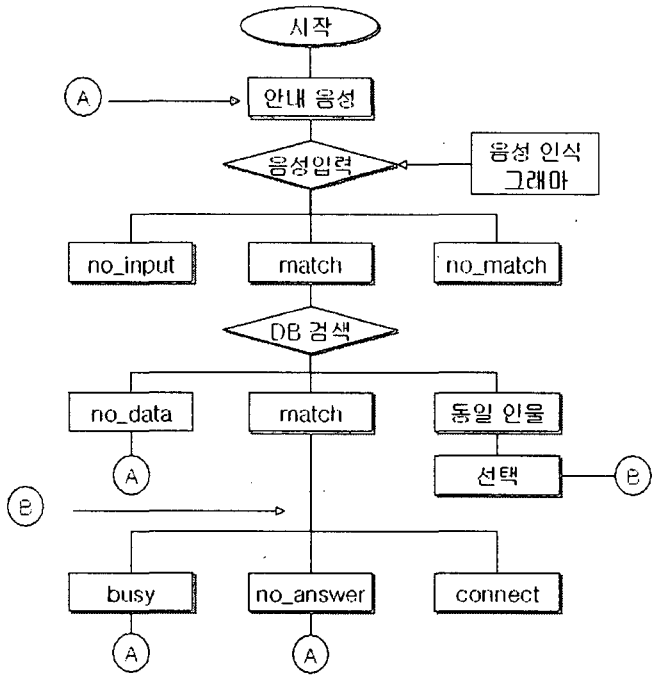
그래마는 음성 인식기에 사용되는 단어 집합 및 단어간의 연관관계를 정의한다. VoiceXML은 음성 입력을 받아들이는 부분에서 Boolean, digit 등의 미리 정의된 그래마 타입 혹은 사용할 그래마를 지정하고, 이를 해석하여 음성 인식기에 전달해야 한다. 많은 벤더들은 VoiceXML의 그래마 형식으로 ABNF(Augmented BNF) 및 SRML(Speech Recognition Markup Language)를 사용한다[4]. 위와 같은 그래마 형식은 단어 집합을 Context Free 그래마로 표현하여 다양한 형태의 문장 혹은 단어 집합을 생성할 수 있다. 본 논문에서 사용된 VoiceXML 해석기는 위와 같은 그래마 형식을 지원한다.

4. VAD (Voice Activated Dialing) 서비스

VAD는 사내에서 특정번호를 누른 후 안내 음성 중간의 아무 때나 연결을 원하

는 사원의 이름을 말할 경우 자동으로 사원의 사무실 혹은 휴대폰으로 연결되는 서비스를 말한다. 현재 KT 서비스개발연구소에서는 VoiceXML 기반의 음성인식 시스템인 VAD 서비스를 제공하고 있다. 처음 사용할 때 음성인식의 어색함만 극복한다면 사원의 이름을 아는 것으로 번호를 외우거나 수첩을 찾을 필요가 없으며, 특히 외부에서 급한 용무중일 때 매우 유용하게 쓰인다.

기존의 시스템들은 음성 시나리오가 하드 코딩 혹은 독자적인 언어를 가지고 운용되어 왔으나, 음성 전용 마크업 언어인 VoiceXML을 이용하면 몇몇 VoiceXML 문서만으로 서비스를 운용할 수 있다. 또한 VoiceXML 문서가 시스템으로부터 독립적으로 존재하고 있기 때문에, VoiceXML 문서를 수정함으로써 안내 음성의 수정과 흐름 변경 할 수 있다. VAD의 VoiceXML 문서의 흐름은 그림 6과 같다.



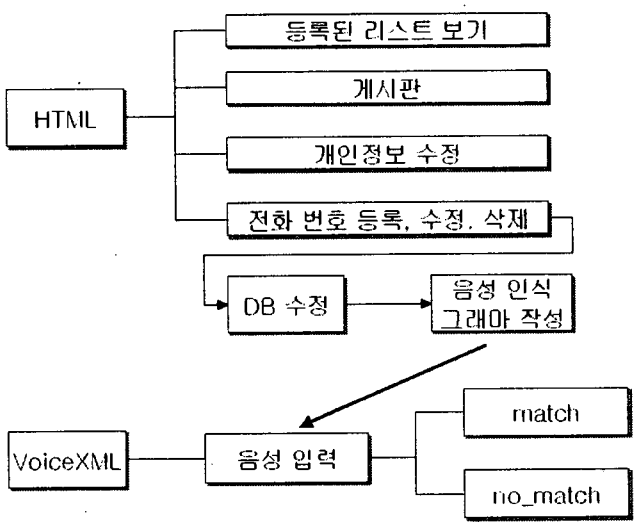
<그림 6> VAD 시나리오의 구성

그림 6과 같이 전화를 걸면 먼저 "연결을 원하는 분의 설명을 말씀해 주세요"라는 안내 음성이 나온다. 이때 연결을 원하는 이름을 음성으로 말하게 되면 음성인식 그래마에서 정의된 이름을 찾고 만약 해당 이름이 없거나 음성입력이 없을 경우 다시 처음으로 돌아간다. 음성 인식 그래마에서 일치되는 이름을 찾았을 경우, 데이터베이스를 통해 대상 전화번호를 검색한다. 데이터베이스에 전화번호가 없을 경우 다시 재입력을 받기 위해 처음으로 돌아간다. 만약 동일 인물이 있을

경우 동일인물에 대한 상세 정보를 주어 사용자가 선택할 수 있도록 한다. 이 서비스에서는 대부분 인명을 인식하도록 되어 있으며 통신실과 같이 직원들이 공통적으로 이용할 수 있는 부서만 부서명을 인식할 수 있도록 하고 있다. 또한 같은 본부 산하의 기관이라도 지역적으로 떨어져 있는 경우에도 연결을 해 주고 있으며 단어인식을 포함해서 문장인식이 가능하도록 설정되어 있다. 그래서 "홍길동", "홍길동 과장님", "홍길동 과장님 부탁합니다." 등으로 사용할 수 있다. 또한 핸드폰 연결도 가능해서 "홍길동 핸드폰 부탁합니다."로 이용이 가능하다. 이와 같은 서비스 시나리오는 VoiceXML 문서로 간단히 표현 가능하다.

5. 개인전화번호 관리 서비스

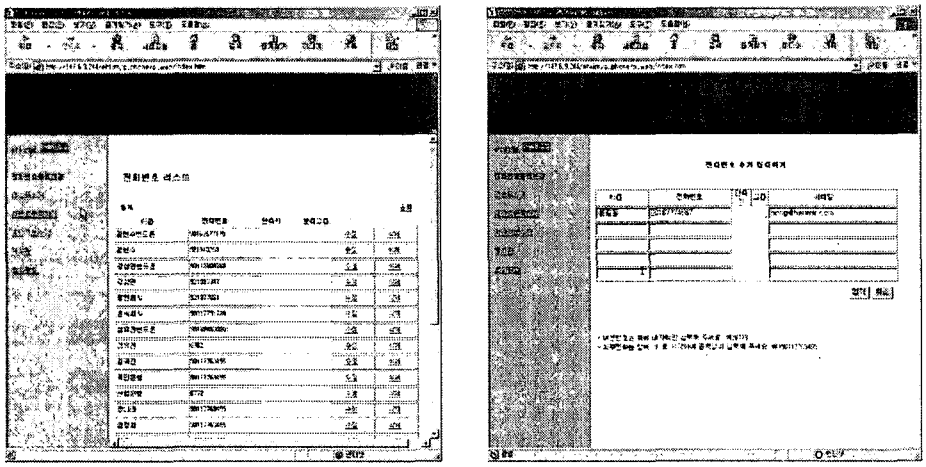
개인전화번호 관리 서비스는 개인 전화번호 수첩과 같이 수많은 사람들의 전화번호들을 웹에서 관리한다. 전화 번호의 추가, 삭제, 수정 등은 일반적인 HTML 기반의 페이지를 통해 수행하고, 자료 검색은 전화를 통해 등록된 사람의 이름을 말함으로써 이루어진다. 즉, 기존의 HTML 서비스가 가지고 있는 자료의 추가, 삭제 등의 편리함과, 전화를 통해 원하는 사람에게 자동으로 연결해줄 수 있는 환경을 제공하는 VoiceXML 솔루션을 접목한 것이다. 그림 7은 이러한 구성을 간략히 보여준다.



<그림 7> 개인전화번호 관리 서비스의 구성

그림 8의 (a), (b)와 같이 HTML 페이지를 통해 새로운 인식명칭을 추가, 삭제,

수정 및 관리를 한다. 새로운 인식명칭이 추가될 경우 데이터베이스를 수정하고, 음성 인식 그래마를 수정한다. 이러한 음성 인식 그래마는 VoiceXML 문서에서 사용되는 것으로써, 문서 표준으로 정의된 ABNF를 통해 기술된다.



(a) 전화번호 리스트 (b) 전화번호 등록

<그림 8> 개인전화번호 관리를 위한 WebPage 화면

이와 같이 작성된 음성인식 그래마를 이용하여 개인에게 특화된 VAD 서비스를 제공한다. 위와 같이 기존의 HTML 기반 시스템에 VoiceXML기반 음성인식시스템을 확장할 때는 VoiceXML 문서와 기존의 HTML 문서간에 데이터베이스와 음성인식 그래마를 공유하여 구성한다.

6. 결 론

대화형 음성언어 인터페이스는 음성으로 인터넷 서비스를 제공하는 음성포탈 서비스를 말한다. 유무선 전화망 혹은 PC 환경을 통해 인터넷 서비스를 제공하는 대화형 음성언어 인터페이스의 필요성이 증가하고 있다. 또한, 음성언어 인터페이스 기술은 21세기를 선도할 수 있는 10대 기술 중의 하나로서, 음성은 영상과 더불어 사람과 기계의 정합 (MMI: Man Machine Interface)방법으로 쉽고 빠르게 정보를 주고받을 수 있다. 이와 같이 VoiceXML 기반한 음성서비스는 장비 제공자 및 정보 제공자에게 새로운 비즈니스 모델을 제시할 수 있고 사용자들에게는 인터넷의 풍부한 지식을 전하는데 큰 의의를 가진다.

본 논문은 이러한 요구에 맞추어 VAD서비스와 개인전화번호 서비스를 VoiceXML

기반으로 개발하였다. 이와 같이 VoiceXML 관련 기술을 이용하여 음성 정보 서비스를 쉽게 제공할 수 있으며, 실시간으로 변하는 정보를 쉽게 음성화할 수 있는 장점이 있다.

참 고 문 헌

[1] Goose, S., M. Newman, C. Schmidt and L. Hue (2000), Enhancing Web accessibility via the Vox portal and a Web-hosted dynamic HTML VoxML converter, *Computer Networks* Vol.33, pp.583~592.

[2] VoiceXML Forum (2001), VoiceXML (Voice eXtensible Markup Language) Ver.2.0.

[3] Koo, M. W. (1999), An utterance Verification system based on subword modeling for a vocabulary independent speech recognition system, In *Proc. European Cont. on Speech Communication and Technology,* pp.287~290.

[4] Speech Recognition Grammar Specification (2001), http://www.w3.org/TR/speech grammar/.

[5] Ferencz, A., S. Choi, H. Song and M. Koo (2001), Hansori2001-Corpus-based Implementation of the Korean Hansori Text-to-Speech Synthesizer, In *Proc. European Cont. on Speech Communication and Technology,* pp.841~844.

[6] World Wide Web Consortium (1998), Document Object Model (DOM) Level 1 Specification.

접수일자: 2002년 5월 8일
게재결정: 2002년 5월 24일

▷ 김학균(Hak-Gyoon Kim)
주소: 137-792 서울시 서초구 우면동 17
소속: KT연구개발본부 서비스개발연구소 음성인식서비스개발팀
전화: 02)526-6772
Fax: 02)526-5909
E-mail: playbug@kt.co.kr

▷ 김은향(Eun-Hyang Kim)
주소: 137-792 서울시 서초구 우면동 17
소속: KT연구개발본부 서비스개발연구소 음성인식서비스개발팀
전화: 02)526-6773
Fax: 02)526-5909
E-mail: hyangii@yahoo.co.kr

▷ 김재인(Jae-In Kim)
주소: 137-792 서울시 서초구 우면동 17
소속: KT연구개발본부 서비스개발연구소 음성인식서비스개발팀
전화: 02)526-5093
Fax: 02)526-5909
E-mail: jaeinkim@kt.co.kr

▷ 구명완(Myoung-Wan Koo)
주소: 137-792 서울시 서초구 우면동 17
소속: KT연구개발본부 서비스개발연구소 음성인식서비스개발팀
전화: 02)526-5090
Fax: 02)526-5909
E-mail: mwkoo@kt.co.kr