모바일 환경에서 VoiceXML 기반의 VUI 개발에 관한 연구

임채욱*, 장민석
국립군산대학교 컴퓨터정보과학과
e-mail : whitewing@kunsan.ac.kr*, msjang@kunsan.ac.kr

Study on Development of VUI Based on VoiceXML in Mobile Environment

Chae-Uk Lim*, MinSeok Jang
Dept. of Computer Information Science, Kunsan National University

요 약

기존의 모바일 디바이스(휴대전화, PDA 등)의 인터페이스는 GUI 방식이 주류를 이루고 있으며, 약간의 음성인식 기술이 접목되고 있는 실정이다. 그 음성인식 기술의 활용은 음성인식 다이얼링에 제한되어 있는 실정이다. 이러한 한계점을 극복하기 위해 본 논문에서는 VoiceXML 포럼에서 제안한 VoiceXML 버전 2.0 스펙을 따르는 VoiceXML 을 모바일 환경에 적용시켜 음성인식 다이얼링 기능 뿐만 아니라, 음성인식 및 합성 기술을 이용한 메뉴선택, 정보 청취 등의 기능을 가능하게 하는 목적으로 VoiceXML 기반의 VUI(Voice User Interface) 개발을 위한 요구사항을 제시하고자 한다. 이는 기존의 GUI 방식뿐만 아니라 VUI 방식을 수용하게 함으로써 사용자들에게 인간친화적인 정보획득 환경을 제공할 것이다.

1. 서론

현재의 모바일 환경에서는 GUI 방식의 인터페이스가 그 주류를 이루고 있다. 하지만 이는 다음과 같은 결정적인 단점을 가지고 있다. 입력 장치로 스타일러스와 터치패드가 있지만 작은 화면을 큰 손을 움직여 가며 클릭하기란 여간 힘든 것이 아니다. 필기인식이라는 방법도 있지만 정보입력의 수단으로는 매우 불편하며, 간단한 것은 SIP(Supplementary Input Panel)을 이용해서 찍는 것이 더 빠를 때가 많다. 이렇게 모바일 디바이스이기 때문에 발생하는 근본적인 단점을 해결하기 위해 본 논문에서는 그 장치에 VoiceXML 을 적용시킨 VUI 방식의 인터페이스를 제공하기 위한 방법을 제안하고자 한다. 현재 VoiceXML 은 VoiceXML 포럼에서 AT&T, IBM, 루슨트 테크놀로지, 모토롤라 등 정보통신 분야의 4 대 기업체를 중심으로 표준화가 진행 중이며 2003 년 현재 W3C 에 Draft 되어있다[1][2].

2. 음성인식 및 합성

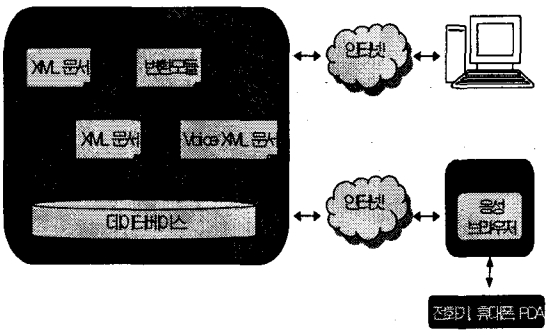
음성인식이란 음성이 담고 있는 정보를 추출해 컴퓨터를 통해 음성의 의미를 알아 내는 기술이다. 즉 인간의 음성을 컴퓨터가 분석해서 인식 및 이해하는 기술을 말하며, 이는 크게 세 가지 측면에서 분류되는데 먼저 적용화자에 의한 분류로는 화자 종속 음성인식 기술과 화자 독립 음성인식 기술이 있으며, 발음방식에 의한 분류로는 고립단어 음성인식 기술, 연결단어 음성인식 기술, 핵심어 인식 기술, 연속 음성인식 기술 그리고 대화 음성인식 등 5 가지 기술로 분류된다. 마지막으로 인식 어휘에 따른 분류로는 소용량 어휘 인식 기술, 대용량 어휘 인식 기술, 가변 어휘 인식 기술 그리고 어휘 독립 인식 기술이 있다.

음성합성(TTS: Text-To-Speech)이란 문자 정보 또는 기호를 인간의 음성으로 변환하여 들려주는 기술이다. 그 방법은 언어의 모든 음소에 대한 발음 데이터베이스를 구축하고 이를 연결시켜 연속된 음성을 생성하게 되는데, 이 때 음성의 크기, 길이, 높낮이 등을 조절해 자연스러운 음성을 합성해낸다. 이를 위해 자연

어 처리 기술이 포함되는데 이러한 음성합성의 분류 체계에 대해 살펴보면 다음과 같다. 제한된 어휘 음성 합성 기술과 무제한 어휘 음성합성 기술이 있다. 전자는 필요한 음성 조각을 미리 녹음했다가 이를 연결시켜 연속된 음성을 만들어내는 기술로 우리 주위에서 쉽게 볼 수 있는 자동응답장치(ARS)나 114 의 전화번호 안내 등이 있으며, 후자로는 자연스러운 음성을 위해서 실제 인간의 음성과 유사하게 구현한다.[6]

3. VoiceXML 기술 및 구조

VoiceXML 은 음성을 통한 대화(Dialog)를 정의하는 언어로서 전화기를 단말로 사용하여 기존의 웹에서 제공하는 정보들을 음성을 통해 브라우징하기 위해 고안되다. 특히 멘트파일 송출과 DTMF 입력처리 뿐만 아니라, TTS 를 통한 실시간 데이터의 음성출력과 ASR(Automatic Speech Recognition)을 통한 사용자의 음성을 인식하여 정보의 흐름을 제어함으로써 보다 자연스럽고 쉬운 정보의 접근을 가능하게 한다. 이는 표준적인 언어를 사용하여 기존에 구축된 인터넷 환경을 활용하게 함으로써 새로운 어플리케이션의 개발과 유지가 보다 쉽고 빠르게 이루어질 수 있게 한다. VoiceXML 기반 시스템의 전체적인 구성을 이해하기 위하여 아래 [그림 1]을 보자.



[그림 1] VoiceXML 기반의 웹 ARS 시스템 통합구성도

그림에서 보듯이 XML 문서, VoiceXML 문서, 변환 모듈, 데이터베이스 등으로 이루어져 있다. 구성에 따르면 각 모듈의 동작을 통해 인터넷 망에 접속된 컴퓨터 뿐만 아니라 전화기나 휴대전화, PDA 등을 통해서도 정보를 얻을 수 있다.

VoiceXML 은 음성 입출력 기반의 음성서비스 시나리오의 표준으로 XML 에 기초하며 다이얼로그, 문법, 이벤트, 오디오 입출력, 콜 제어, 흐름 제어 등에 관련된 엘리먼트들로 구성되며[3], VoiceXML 을 모바일 환경에 접목하기 위해 사용할 수 있는 VoiceXML 엘리먼트들은 아래 [표 1]과 같다.

[표 1] VoiceXML 의 Element 추출

분류항목	상세 내용
이벤트	noinput, nomatch, throw, catch, error, help, filled
조건문	if, elseif, else, option
음성관련	block, break, prompt, voice, sentence, grammar
Connect	link
변수관련	var, value
분기문	goto

위의 [표 1]은 VoiceXML 버전 2.0 에 있는 엘리먼트 중에서 본 논문에서 사용할 엘리먼트를 추출하여 VUI 시나리오에 사용한다. 모바일 환경 중 PDA 환경에서 구현하기 위해서 DTMF 등 전화기 특성을 고려한 엘리먼트와 실제 시나리오에 불필요한 엘리먼트들을 제외하였다.

4. 모바일 환경의 단점과 극복방안

본 장에서는 모바일 환경에서의 VUI 를 개발하기 위해, 모바일 환경과 유선환경의 차이를 비교하고 VoiceXML 을 모바일 환경에 적용할 때 발생할 수 있는 문제점을 도출하며, 이를 해결할 수 있는 방법을 모색하고자 한다.

모바일 환경과 유선 환경은 하드웨어와 소프트웨어 측면에서 비교할 수 있다. 전자 측면에서는 모바일 환경은 유선 환경에 비해서 메모리가 많이 부족하여 대용량의 프로그램을 설치할 수 없고, 액정 크기가 작아서 많은 정보를 한꺼번에 볼 수 없다. 그리고 입력 장치가 유선 환경 즉, 키보드보다 불편하여 사용자들이 문자 등을 입력할 때 많은 불편을 느끼게 된다. 후자 측면에서는 메모리가 부족해서 대용량의 프로그램을 설치할 수 없기 때문에 운영체제 및 일반 어플리케이션도 유선 환경보다는 용량이 작은 운영체제나 어플리케이션을 사용하게 된다. 그 대표적인 예로 Windows CE, MS SQL Server CE 등을 들 수 있다. 이러한 모바일 환경에서 음성인식 기능을 수행하려면 유선 환경과는 다른 점들을 고려해야 한다. 아래 [표 2]는 모바일 환경과 유선 환경의 차이점을 보여주고 있다.

[표 2] 모바일 환경과 유선 환경의 차이점

	모바일 환경	유선 환경
메모리 (RAM/ROM)	최소 200KB	최소 32MB
입력장치	스타일러스, 터치 패드	키보드, 마우스
운영체제	Windows CE 등	Windows 2000 등
액정크기	3.77 인치	14 인치 이상
해상도	240 X 320	800X600 / 1024X768

[표 2]에서 보다시피 기존의 유선 환경에서의 VUI 와 모바일 환경에서 VUI 는 달라질 수 밖에 없다[9]. 기본적으로 위 문제점을 해결하기 위해서는 첫 번째로 기존의 유선 환경에서 사용하던 언어가 아닌 임베

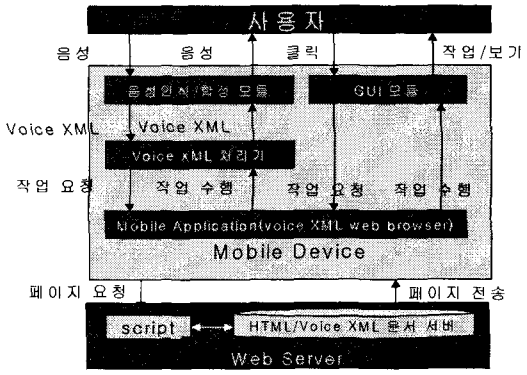
디드 언어를 사용하여야 하며, 두 번째로 현재 입력 장치인 스타일러스나 터치패드가 아닌 사람의 음성을 통해서 어플리케이션을 수행할 수 있는 음성 인터페이스를 제공해야 한다. 그리고 또 다른 문제점은 모바일 환경에서는 유선 환경보다 잡음의 영향을 많이 받게 된다. 그 잡음을 어느 정도 제거하기 위해서 특징 파라미터를 추출하는 과정을 거쳐야 하는데 음성인식을 위하여 주로 사용되는 특징은 LPCC, PLP, MFCC 등이 있고, 이들을 이용해서 잡음을 제거하게 되는데 본 논문에서는 MFCC 를 이용하여 잡음에 민감한 고차 영역을 감소시킴으로써 잡음을 제거하고자 한다. 아래 [그림 2]는 MFF 에서의 특징 파라미터를 추출하는 과정이다.

[그림 2] MFCC 에서의 특징 파라미터 추출 과정

본 논문에서는 전처리 과정과 윈도우는 따로 처리해줘도 되지만 같이 처리하도록 하겠다. 전처리 과정과 윈도우를 거치지 않고 바로 FFT 과정으로 넘어가게 되면 주파수 성분만 나타나고, 시간적인 정보는 하나도 나오지 않게 된다. 그러나 음성에서는 시간적 정보도 있어야 하기 때문에 전처리 과정과 윈도우에서 이를 반영해주어야 한다. FFT(Fast Fourier Transform)는 신호처리에서 시간 영역의 신호를 주파수 영역으로 변환시키는데 많이 사용되고 있다. 앞에서 구한 필터 뱅크의 출력 에너지를 그냥 사용하는 것이 아니라 로그를 취하게 되는데 그 이유는 우리의 귀가 소리의 크기에 대해 로그 함수로 느끼기 때문이다. DCT(Discrete Consine Transform)을 마지막으로 해주면 멜켑스트럼을 얻을 수 있다. DCT 는 필터 뱅크의 출력 간의 상관 관계를 없애주고 파라미터의 특징을 모아주는 역할을 한다[4][7].

5. 모바일 환경에서의 VUI 설계 구성 및 인식 알고리즘

본 논문에서는 우리가 실생활에서 흔히 사용하는 모바일 디바이스인 PDA 에서의 VUI 를 설계하고자 한다. [그림 3]은 궁극적인 모바일 환경에서의 VUI 의 구성도이다. 사용자가 모바일 디바이스에 음성으로 정보를 입력하면 모바일 디바이스는 그것을 음성인식/합성 모듈을 통하여 인식 및 합성하여 VoiceXML 의 시나리오를 통하여 모바일 디바이스 어플리케이션에 전달되게 된다. 그러면 모바일 디바이스 어플리케이션에 작업 요청을 하게 되고 모바일 디바이스 어플리케이션은 작업 수행을 하게 된다.



[그림 3] 궁극적인 모바일 환경의 VUI 구성도

모바일 디바이스가 음성의 의미를 이해하려고 하면 음성 파일을 그냥 들려주면 안 된다. 인식에서는 음성에 포함되어 있는 정보를 최대한 잘 나타낼 수 있는 파라미터를 특징 파라미터로 사용하고 있다. 이러한 특징 파라미터로는 MFCC(Mel-Frequency Cepstral Coefficient)와 LPCC(Linear Predictive Coefficient Coefficient), PLP 등이 있다. 예전에는 LPCC 가 주로 사용되었다. 왜냐하면 계산량이 적고 주파수 영역으로 변환하지 않고도 구할 수 있으며, 성도의 특성을 잘 나타내기 때문에 인식에서 특징 파라미터로 사용되었다. 하지만 요즘은 잡음에 더 강하고 인간의 청각 특성을 고려한 멜켑스트럼(MFCC)과 PLP 특징 파라미터가 주로 이용된다. 특히, MFCC 는 Mel-scale Frequency 영역에서 특징 파라미터를 구하므로 여러 인식 시스템에서 성능이 좋은 것으로 나타났다. 멜켑스트럼이라는 음성인식에서 사용하는 특징 파라미터에 대해서 알아보겠다. 언급했듯이 컴퓨터는 특징 파라미터를 통해서 음성을 구분하게 된다.

인식 알고리즘으로는 DTW(Dynamic Time Warping)와 HMM(Hidden Markov Model), NN(Neural Network) 등이 주로 사용되고 있다. 다들 아다시피 모든 알고리즘에서 인식하고자 하는 기준 음성의 특징을 갖고 있거나, 많은 데이터를 갖고 훈련된 기준 모델을 갖고 있다. *DTW 는 특정 음성의 기준 특징 파라미터와 테스트 음성의* 특징 파라미터의 거리를 구해서 음성을 인식하는 방법으로 비교적 간단히 음성 인식기를 구현할 수 있으며, 기준 모델에 대한 훈련 과정이 필요 없다. 하지만 알고리즘이 간단한 만큼 인식할 수 있는 단어의 수가 적으며 모든 기준 특징 파라미터와 테스트 특징 파라미터를 비교하기 때문에 인식 단어 수가 증가함에 따라 계산량 증가를 가져오게 된다. 한편 HMM 은 훈련 과정을 갖고 있으며, 훈련된 모델과 테스트 음성의 확률적 거리를 구해서 음성인식을 하게 된다. 그러므로 훈련하기 위해 많은 음성 데이터가 필요하게 되며, 훈련 시간도 오래 걸리게 된다. 하지만 DTW 보다 인식 단어 수가 많기에 현재 많이 이용되고 있는 알고리즘이다. 마지막으로 NN 은 인공지능을 이용해 기준 모델과 테스트 데이터를 비교해서 인식하는 방법이다. 이 논문에서는 비교적 구현하기 쉬운

DTW 를 이용해 간단한 단어 인식기를 구현해서 모바일 환경에 접목시킬 것이다[4].

6. 결론 및 향후 연구

본 논문에서는 모바일 환경에서의 VoiceXML 기반의 음성 인터페이스 환경을 제공함으로써 모바일 디바이스를 사용자에게 보다 쉽게 사용할 수 있는 환경을 제공하는 것을 궁극적인 목표로 삼으며, 이를 위해 수행되어야 구체적인 연구의 최종 목표는 기존의 음성처리기술과 VoiceXML 기술을 접목하기 위해 요구되는 요소 기술을 개발하는 것이다.

향후 연구 과제로는 본 논문의 제안 사항을 실제 구현함에 있어서 발생할 수 있는 문제점 극복과 모바일 환경에서의 정확도 문제를 극복할 수 있는 소음이 차단 및 해결 방안에 대한 연구가 요구된다

참고문헌

[1] 예상후, VoiceXML 을 이용한 VUI Web Browser 구현 및 설계, 군산대학교 컴퓨터정보과학과 석사학위 논문, 2003.2

[2] VoiceXML 포럼, http://www.voicexml.org

[3] 이인숙, 홍기형, "유무선 전화 음성 기반 VoiceXML 학습 평가 시스템", 한국멀티미디어학회지, 제 5 권 제 4 호, p62~p63, 2001 년 12 월

[4] 마이크로 소프트웨어 잡지, 2000 년 9 월호 (p246~p247), 2002 년 11 월호(p410~p411), 12 월호 (p422~p425), 2003 년 1 월호(p370~p371).

[5] 김경란, VoiceXML 기반 음성 브라우저의 설계 및 구현, 성신여자대학교 전산학과 석사학위 논문, 2001

[6] 휴렛패커드, http://www.hp.co.kr

[7] http://inc2.ucsd.edu/~owkwon/