2001년 한국정보처리학회 춘계 학술발표논문집 제8권 제1호

VXML을 이용한 개인 전화 도우미의 설계 및 구현

°하준, 윤영선, 은성배
한남대학교 정보통신공학과
e-mail:{jha, ysyun, sbeun}@daniel.hannam..ac.kr

Design and Implementation of Personal Telephone Assistant using VXML

Jun Ha, Young-Sun Yun, Seongbae Eun
Dept. of Information and Communication Eng., Hannam University

요약

VXML은 사람 목소리나 전화기의 톤과 같은 음향 입력과 컴퓨터에 의해 합성 또는 녹음된 목소리를 들려주는 음성 출력을 가지는 음성 브라우저를 위한 웹 저작 언어이다. 본 논문에서는 VXML을 이용하여 작은 규모의 회사나 SOHO 사업자들에게 최적화된 ARS 기능, 음성 인식 기술을 이용하여 음성 메시지 저장 및 지능적 검색 기능, 사용자 부재 시 착신 전환 또는 SMS 통보를 할 수 있는 개인용 전화 도우미 (PTA; Personal Telephone Assistant)의 설계 및 구현에 관하여 기술한다. PTA는 VXML 인터프리터를 내장하고 동적으로 VXML 문서를 적재함으로써 고객맞춤형의 ARS 기능을 지원한다는 장점을 갖는다.

1. 서론

사람과 컴퓨터의 접속 방식 (Human-Computer Interface)으로 음성이 각광을 받기 시작하면서, 데스크탑 컴퓨터로만 할 수 있었던 많은 업무들이 이동형 컴퓨터나 특수 목적의 컴퓨터, 전화 등을 통하여 처리할 수 있게 되었다. 인터넷이 발달하면서 서비스의 대부분을 차지하는 WWW(World Wide Web)가 그 예라 할 수 있다. WWW은 HTML(Hyper Text Markup Language)을 이용하여 하이퍼텍스트 형태로 탐색한 자료를 검색하고 보여주는 역할을 하는 서비스이며, 무엇보다도 텍스트뿐만 아니라 그림, 음성, 애니메이션과 같은 다양한 미디어를 제공하기 때문에 폭발적인 인기를 끌고 있는 서비스이다. 이러한 서비스를 음성을 이용한 입·출력 방식에 이용하기 위해서는 사람과 컴퓨터간의 별도의 접속 방식이 필요하게 된다. 이러한 목적으로 1999년에 AT&T, IBM, Lucent Technology, Motorola 등 정보 통신 분야의 기업들에 의해 대화형 마크업 언어 (dialogue markup language)인 VXML(Voice eXtensible Markup Language)이 제안되었다.

VXML은 HTML과 같이 사람과 컴퓨터간의 대화를 표현하기 위한 웹 기반의 마크업 언어라고 정의할 수 있다. HTML이 키보드와 마우스 등의 입력 장치와 디스플레이와 같은 출력 장치를 이용한 그래픽 웹 브라우저를 가정한다면, VXML은 사람 목소리나 전화기의 톤과 같은 음향 입력과 컴퓨터에 의해 합성 또는 녹음된 목소리를 들려주는 음성 출력을 가지는 음성 브라우저를 가정하고 있다. 따라서 컴퓨터의 접근이 어려운 장소에서 전화 등의 음성 입·출력 장치를 이용하여 인터넷의 활용과, 음성을 이용한 컴퓨터의 접속 관계를 체계적으로 정리하여 다양한 서비스가 가능해졌다.

일반적으로 VXML을 이용한 서비스는 널리 사용되고 있는 전화나 음성 입·출력 장치를 갖는 컴퓨터 환경에서 이용되고 있다. 일반 전화를 통한 VXML 서비스는 인터넷과 전화 회선 망 (PSDN)을 연결해주는 음성 게이트웨이를 이용하여 제공되고 있으며, 그 범위는 나날이 확대되고 있다[1].

전화를 통하여 사람의 음성으로 인터넷에 접속하거나 다른 서비스를 받을 수 있다는 점 때문에 많은 업체에서 VXML을 이용한 응용 서비스를 개발하고 있다. 이중 대표적인 예로 정보 검색 (Information retrieval)을 들 수 있다. 이 서비스를 이용하면 운전을 하거나 이동 중에 교통 상황이나 뉴스, 주식 정보, 기상 예측 등과 같은 정보를 쉽게 검색할 수 있다. 다음으로는 전자 상거래가 있다. 전화를 통하여 상품을 주문하는 경우와 같이, 교환원을 통하지 않고 곧바로 상거래 시스템에 접속하여 구매하고 배달을 지시할 수 있으며, 계좌 조회, 증권 거래와 같은 은행 업무를 할 수 있다. 또한 음성으로 전화를 걸거나 부재중 걸려온 전화를 대신 받고 메시지를 남기는 전화 비서 등의 업무를 대신할 수 있다. 마지막으로 전화와 컴퓨터를 연결하여 수신된 전자 메일을 검색하여 읽어주거나 음성이나 전자 메일을 대신 보내는 통합 메

시징 서비스(UMS; Unified Messaging System)들을 생각할 수 있다. 이외에도 다양한 방식의 서비스들이 꾸준히 개발되고 있다.

본 연구에서는 이와 같이 다양한 서비스를 제공할 수 있는 VXML을 이용하여 작은 규모의 회사나 SOHO 사업자들에게 VXML을 이용하여 쉽게 구축 가능한 간단한 ARS 기능과, 음성 인식 기술을 이용하여 음성 메시지 저장 및 지능적 검색 기능, 사용자 부재 시 착신 전환 또는 SMS 통보를 할 수 있는 개인용 전화 도우미 (PTA; Personal Telephone Assistant)의 설계를 제안하고 간단한 프로토타입을 선보인다. 이와 같은 기능은 지금까지 전화 교환기 시스템에서 담당해왔기 때문에 그 비용이 높고 사용법이 어렵다는 단점이 존재한다. 따라서, 본 연구에서는 전화 교환기를 거치지 않고 전화기 안에서 VXML과 음성 인식 시스템을 이용하여 전화 교환기가 제공하였던 여러 기능을 저렴한 가격으로 구현할 수 있는 전화 도우미를 제안한다. 제안된 시스템은 인터넷에 연결된 컴퓨터와 같이 사용되면 수신된 전자 우편이나 음성 메일을 들려주거나 대신 전송하는 UMS 기능도 제공하도록 설계되었다.

2. 관련 기술 및 연구

2.1 VXML (Voice Extensible Markup Language)

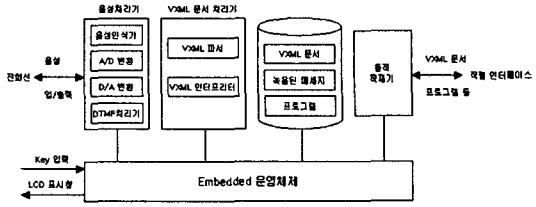
VXML은 AT&T, IBM, Lucent Technology, 모토롤라 등이 연합하여 결성한 Voice XML Forum에서 제안하였다. 1999년 말 Spec. 0.9가 발표되었고 2000년에 Spec. 1.0이 발표되었다. 현재 VXML 포럼을 중심으로 활발한 연구가 진행중이며 몇몇 국내 기업에서도 VXML 시스템 개발 및 상품화에 주력하고 있다.

VXML은 XML(Extensible Markup Language)을 기반으로 음성 인터페이스를 구현하기 위한 표준이다. 주요 기능으로 음성의 저장 및 출력, 음성 인식, DTMF 인식 등을 정의하고 있으며, 이를 이용한 Web 기반의 대화형 음성 응답 시스템의 개발을 목표로 한다. VXML을 이용한 대화형 음성 응답 시스템은 PC, 노트북 등의 인터넷 접속 장치 없이 전화, PCS 등의 음성 기기만으로 정보의 전달, 저장 및 Web 검색 등의 기능을 제공한다. 이러한 시스템을 구현하기 위해서는 VXML로 만들어진 문서를 해석, 수행하는 VXML Interpreter를 구현해야 하며 이 Interpreter를 중심으로 음성 인식 시스템, TTS(Text to Speech) 시스템, DTMF 인식 시스템, Web Server, DBMS(Data Base Management System) 들을 유기적으로 통합하는 기술이 필요하다.

2.2 음성 인식 시스템 (Speech Recognition System)

음성 인식이란 "음성에 포함된 언어적인 정보를 추출하여 컴퓨터가 해독할 수 있는 표현 방식으로 변환하는 과정"을 말하는 것으로 음향학·음운학·언어학 등 단계적인 처리과정을 필요로 한다. 효과적인 음성 인식을 위해서는 다양한 사항들이 고려되어야 하는데, 특히 화자의 발성 방법, 화자 의존성, 사용 어휘, 인식 단위, 문법, 환경적인 요소들을 들 수 있다. 이러한 요소들은 인식 성능 및 복잡도와 밀접한 관계를 가지므로 적용할 대상에 따라 적절한 조합을 선택하여야 한다.

(그림 1) PTA의 구조



일반적인 음성 인식 시스템은 마이크나 전화로 입력된 음성을 단 구간 (short time)별로 분할하여 특징 분석을 통하여, 음성학적 특징을 잘 표현하는 음성 특징 계수들로 나타낸다. 이렇게 추출된 특징 벡터는 미리 저장된 음소 또는 단어 단위의 모델을 이용하여 패턴 인식과정에서 가장 적합한 후보 음소 또는 단어 열을 생성하게 된다. 마지막으로 언어 처리부에서는 후보 음소·단어 열들의 정보를 토대로 인식 대상 어휘 및 문법 구조, 또는 특정 주제의 부합 여부를 판단하여 최종 인식된 문장을 출력하게 된다[2].

인식 단위나 어휘의 양, 그리고 발성 방법에 따라 음성 인식 기술을 구분할 수 있는데, 일반적으로 분류하는 대표적인 방법이 발성 방법이다. 즉, 인식 대상 어휘를 한 단어씩 또박 또박 발성하느냐, 또는 단어들을 연결하여 발음하거나, 연속적으로 발음을 하느냐에 따라 고립 단어 인식, 연결 단어 인식, 연속 음성 인식으로 분류할 수 있다. 각 음성 인식 시스템은 인식률이나 인식 속도를 최적화하기 위하여 인식 단위를 단어 또는 음소 단위로 모델링하고 언어 정보를 사용하기도 한다. 본 연구에서는 인식 대상 방법을 한 단어씩 또박 또박 발성하거나, 자연스럽게 말한 정보에서 중요한 단어 또는 명령어를 선택하여 인식하는 방법을 채택하였다.

3. 설계 및 구현

본 절에서는 PTA의 설계 및 구현에 대하여 기술한다. PTA는 개인이나 작은 사무실 등에서 사용되는 것을 가정하고 있다. 먼저, PTA의 기능을 예를 들어 설명하고 설계시 고려사항을 분석하며 시스템 구성을 설명한다. 그리고 시스템의 주요 부분인 VXML 인터프리터의 구현과 음성인식기에 관하여 설명한다.

3.1. 요구 분석

사무원이 외출하여 사무실에 아무도 없을 때 PTA는 큰 도움이 된다. PTA의 동작을 다음과 같은 시나리오로서 설명할 수 있다.

1) 사무실에 전화가 오면 먼저 미리 녹음된 부재중 메시지를 전달한다

2) VXML로 미리 프로그램된 방식에 따라 사용자의 메시지를 녹음한다. 이때 발신자의 이름, 전화번호, 용건등을 분리해서 녹

음한다.

3) 호출자가 원하는 경우에 원하는 사람에게 전화를 걸고 호를 전환해 준다.

4) 지정된 시간에 지정된 상대방에 전화를 걸어 지정된 VXML 프로그램에 의해 일을 처리한다.

1), 2)의 경우는 기존의 자동응답 전화기와 같은 기능이다. 이때 좀 더 보강된 기능으로는 VXML로 프로그램된 방식에 따라 메시지 내용을 발신자의 이름, 전화번호, 용건 등으로 분리해서 녹음할 수 있다는 점이다. 이러한 기능은 ARS 기능을 이용하면 쉽게 처리할 수 있으나 저가격의 시스템에서 ARS 기능을 사용자의 요구에 맞게 지원하는 것이 어렵다. PTA에서는 이 ARS 기능을 VXML을 이용하여 쉽게 구현한다는 것이 장점이다. 3)의 경우는 기존의 교환기에서 지원하는 착신 전환에 해당한다. 다른 점은 PTA에서는 언제나 착신전환이 일어나는 것이 아니라 몇가지 대화를 통하여 발신자가 원할 때에만 착신전화이 된다는 점이다. 수신자가 원하지 않을 경우엔 전환이 거부되도록 할 수도 있다. 4)의 경우는 지정된 시간에 지정된 번호로 전화하여 미리 녹음된 내용을 전달하는 기능이다.

이와 같이 PTA의 요구 기능을 한마디로 정의하자면 사용자의 용도에 따라 고객 맞춤형의 ARS 기능을 지원한다는 것인데 VXML 인터프리터를 내장하여 필요시 적절한 VXML문서로 교체함으로써 이를 지원한다.

3.2. 설계시 고려사항

PTA는 개인이나 작은 사무실에서 기존의 전화를 대체하는 형상을 갖는다. 설계시 고려사항은 다음과 같다.

1) 기존의 전화기와 같은 형상을 갖는다.

2) 저가격이어야 하며 소형이어야 한다.

3) LCD 표시창, 숫자 판등을 출력 장치로 가지며 PC등과의 통신을 위하여 직렬 통신 채널을 갖는다.

4) 음성인식기나 VXML문서, 미리 녹음된 음성 메시지 등을 쉽게 교체할 수 있어야 한다.

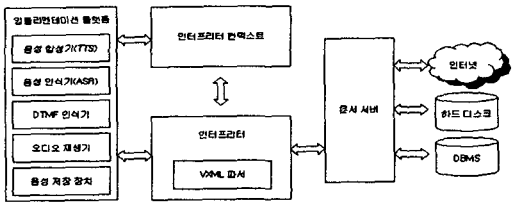
3.3. 전체적인 구조

그림 2 에서 볼 수 있는 것처럼 PTA는 크게 5 부분으로 나뉜다. 먼저 음성 입력 및 출력을 처리하는 음성처리기, VXML 문서를 처리하여 지정된 기능을 수행하는 VXML 처리기, 메모리 기반의 파일시스템, 프로그램이나 VXML문서를 동적으로 적재하는 동적 적재기, 그리고 기본적인 입/출력과 시스템 전체를 통괄하는 운영체제이다.

음성처리기는 A/D 변환 및 D/A변환을 처리하고 전화선을 통하여 입력된 음성을 인식하는 역할을 수행한다. 또한 DTMF 신호를 인식하는 기능도 수행한다. 일반적인 VXML 처리에는 음성합성기를 포함하는데 PTA가 소형이므로 음성을 녹음하여 출력하도록 하였다.

VXML처리기는 정의된 VXML문서를 해석하여 지정된 명령을 처리하는 모듈이다. 음성인식기나 DTMF 처리기로부터 해석된

(그림 2) VXML 시스템의 구조



이벤트에 따라 다양한 기능을 처리하며 미리 녹음된 음성을 출력한다.

메모리 파일 시스템은 VXML 문서나 녹음된 음성 메시지, 프로그램등을 저장하고 검색하는 역할을 수행한다. 플래시메모리를 이용하여 전원이 없더라도 내용을 잃어버리지 않도록 하였다.

동적 적재기는 PTA의 고객맞춤 기능을 위하여 필수적인 기능으로서 직렬 포트로부터 프로그램이나 VXML 문서, 또는 녹음된 음성메시지등을 동적으로 다운로드하는 기능을 수행한다. 직렬포트는 PC와의 연결을 위한 것이며 사용자는 PC를 통하여 인터넷으로부터 다운로드받거나 또는 도구를 수행하여 직접 VXML문서를 생성할 수도 있다.

3.4. 구현

PTA의 구현은 두 단계를 거쳐야 한다. 먼저 PC상에서 각 기능을 구현한 프로토타입 단계이고 다음은 이를 소형의 임베디드 시스템에 이식하는 단계이다. 현재, PC상에서 VXML 인터프리터와 음성인식기 등이 구현 완료되어 통합된 상태이다.

구현될 타겟 시스템은 주 프로세서로서 LOSA(Linux On SA110)를 사용할 예정이다. LOSA[3]는 임베디드 리눅스 운영체제와 메모리, 직렬 포트, 이더넷 제어기 등을 내장하고 있으므로 소형의 내장 시스템을 설계할 때 장점을 갖는다. 음성처리기를 위하여 SCENIX[4] 8비트 마이크로 컨트롤러를 사용할 예정이다. SX-52는 시그널 처리를 위한 하드웨어를 내장하고 있지는 않지만 처리속도가 100MIPS정도로 빠르기 때문에 A/D 변환, D/A 변환, DTMF 인식 등을 모두 소프트웨어로 처리할 수 있다.

임베디드 시스템에 상기 모듈들을 이식할 때에는 여러 가지 문제들을 해결해야 한다. 먼저, PC 상에서 구현된 VXML 인터프리터가 MS의 DOM 파서를 이용하여 구현되었고 MFC 라이브러리를 다수 사용하므로 이부분을 직접 구현해 주어야 한다. 또한 음성인식기의 크기가 너무 크지 않아야 한다는 점도 문제점의 하나이다.

3.5. VXML 처리기

VXML 시스템은 음성의 입출력 및 호 처리를 담당하는 임플리멘테이션 플랫폼, 전체 시스템의 특성 제어와 사용자와의 연결관리를 담당하는 VXML 인터프리터 컨텍스트, 문서의 해석 및 전체 대화를 제어하는 VXML 인터프리터, 인터프리터의 문서 요청을 처리하는 문서 서버의 4가지 모듈로 구성된다.

임플리멘테이션 플랫폼은 사용자와의 입출력을 담당한다. 이를

위해 DTMF 인식 시스템, 음성 인식 시스템, TTS, 음성 저장 시스템들을 인터프리터와 인터프리터 컨텍스트의 명령에 따라 제어하는 기능을 가지고 있다. 따라서 사용자가 체감하는 서비스의 품질은 이 플랫폼에 좌우되며 이 플랫폼의 품질은 플랫폼이 제어하는 다른 시스템들의 품질에 달려있다.

VXML 인터프리터 컨텍스트는 전체 시스템의 특성을 제어한다. 플랫폼을 제어하여 음성을 출력하는 TTS의 출력 음성 속도 제어, 음량 제어, 음성 인식기의 감도 제어 등의 기능을 한다. 또한 사용자와의 연결을 관리하는데 이를 세션이라고 하며 초기 호 처리, 사용자 정보 저장, 사용자 연결 상태 감시, History 등이 그 기능이다.

VXML 인터프리터는 VXML 시스템의 중심으로 VXML 문서의 해석과 전체 대화의 제어를 담당한다. 문서 서버에게 문서를 요청하고 문서를 VXML 파서를 이용하여 해석하며 해석된 문서에 따라 임플리멘테이션 플랫폼을 제어하여 사용자와의 상호 작용을 제어한다.

문서 서버는 인터프리터의 문서 요청을 처리한다. 인터프리터가 요청한 문서를 지역 저장 장치(HDD)나 인터넷에서 찾아 인터프리터에게 넘겨주는 것이 주요 기능이다. 이 과정에서 DBMS와의 연동 하여 문서를 처리할 수도 있다.

VXML 인터프리터는 Microsoft Visual C++을 이용하여 구현하였다. VXML 파서는 MS XML DOM 파서[5]를 이용하였다. DOM 파서는 XML 파서로 VXML DTD(Document Type Definition)와 결합하여 VXML 파서로 이용할 수 있다. 구현된 인터프리터의 클래스 구성은 다음과 같다.

- CInterpreter : 인터프리터 주 클래스
- CDocumentElement : FIA 실행 클래스
- CFormItem : Form Item 처리 클래스
- CCommon : 공통 Element처리 클래스
- CPlatForm : 임플리멘테이션 플랫폼 클래스
- CGrammar : 그래머 저장 및 처리 클래스
- CEcma : ECMA Script 처리 클래스

현재까지 구현된 인터프리터는 표준에 제안된 모든 기능이 구현되지는 않았다. 핵심적인 주요 기능만이 구현되었는데, 지금까지 구현된 인터프리터로도 PTA를 구현하는 것이 가능하다. 이후 인터프리터는 계속해서 기능을 확장해 나갈 계획이며 그에 따라 PTA의 기능도 확장될 것이다.

3.6. 음성 인식 시스템의 구현

전화선을 통하여 전달되는 음성은 A/D 변환기를 통하여 8kHz, 8bit, mu-law 또는 A-law로 저장된다. 녹음된 음성은 20ms 단위의 분석 창을 통하여 10ms씩 이동시키며 특징 계수로 표현되며, 신호의 고주파 성분 잡음 등을 제거하기 위해서 저역 필터를 통과시키고, 성대의 스펙트럼 형태를 더욱 정확하게 표현하기 위하여 pre-emphasis 등을 수행한다. 필터를 통과한 음성은 Fourier 변환을 통과하여 사람의 청각 특성을 표현하는 mel 변환을 거친 후, 13차의 MFCC로 표현된다. 음성의 동적 특성을 표현하기 위하여 구해진 13 MFCC의 1차 변환 값을 포함한 총 26차의 특징 벡터가 인식에 사용된다. 전화 음성에는 회선을 통하여 전달되는 과정에서 channel 잡음이 추가되기 때문에, 이러한 잡음을 제거하기 위하여 일반적으로 많이 사용되는 CMS (Cepstral Mean Subtraction) 방법을 이용하였다. 사용된 인식기는 HMM을 이용하여 고립단어를 대상으로 구축되었으며, 대상 어휘가 가변적이기 때문에 삼 음소(triphone) 단위의 모델을 학습시키고 사전 정보에 의하여 단어 모델을 구성하도록 하였다.

4. 결과 및 향후 연구 방향

본 연구에서는 VXML을 이용하는 개인전화도우미(PTA)의 설계 및 구현에 대하여 기술하였다. PTA의 가장 큰 특징은 소규모 저가격이면서도 사용자 맞춤의 ARS 기능을 제공한다는 것이다. VXML 문서를 필요에 따라 수시로 다운로드하여 원하는 형태의 ARS 서비스를 PTA에 설정할 수 있다. PTA에는 VXML 인터프리터와 음성 인식기 모듈 등이 내장돼 있으므로 이를 가능하게 한다.

본 연구에서 구현한 음성 인식 시스템은 녹음된 전화기의 hand-set이나 환경, 그리고 사용 대상 어휘 시스템에 따라 성능이 변하게 되므로, 일관성 있는 평가가 어렵다. 그러나, 실험실 환경에서 음성 인식 시스템을 테스트한 결과 연속 음성에 대하여 95% 이상의 단어 인식률을 보이고 있어 제안한 PTA에 적용이 가능하리라고 예상된다.

현재, PC상의 프로토타입이 모듈별로 구현되어 통합되는 단계에 있으며 타겟 시스템에서 구현하기 위한 이식 작업을 병행하고 있다.

참고 문헌

[1] E.D.Tober, R.Marchand, J.Ferrans, "VoiceXML Tutorials," http://www.voicexml.org/tutorials/intro1.html

[2] 음성처리산업의 국내외 동향 및 발전방안 연구, 산업자원부, 2000. 12

[3] http://www.sxtech.com

[4] http://www.linuxonchip.com

[5] Microsoft XML 2.5 SDK - XML Developer's Guide, http://msdn.microsoft.com/libary/psdk/xmlsdk/xmlp6bho.htm