모바일 환경에 적합한 음성인식기 설계에 관한연구

A Study on the Speech Recognizer Design in Mobile Environments

최승호
(동신대학교 멀티미디어통신공학전공, 교수)

조제황
(동신대학교 전자공학전공, 교수)

목 차

Ⅰ. 서 론

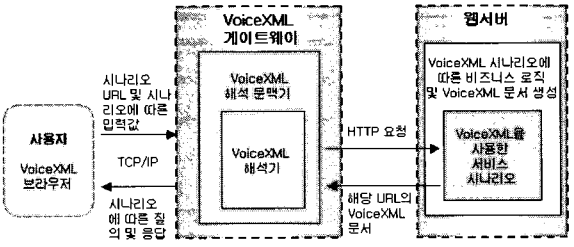
현재 차세대 모바일 인터넷 폰의 등장으로 WAP, IMT2000, 차세대 유무선 통합 서비스들이 대거 출현하면서 무선 환경에서 사용이 편리한 음성인터페이스를 이용한 다양한 서비스들이 출현하고 있다. 이러한 서비스들은 서버중심의 음성인식과 합성 기술을 접목시켜 음성인터페이스를 구현하지만 기존의 서버중심의 음성인식 모듈은 사용자가 급속히 증가하는 유무선 서비스 환경에서, 특히 서버의 빠른 응답시간을 요구하는 무선 환경에서 사용자가 원하는 응답속도를 충족시키지 못할 뿐만 아니라, 음성 데이터 전송량의 과다로 사용자로 하여금 과중한 통신 이용요금의 부담을 안겨 주게 된다. 이러한 무선 환경을 고려해 볼 때, 모바일에 최적화된 음성인식 시스템의 연구가 대두되게 된다. 기존의 유무선 통합 서비스에서 음성 인터페이스를 제공하는 많은 시스템들은 기존의 서비스 환경에서 음성인식 모듈을 접목한 형태로서, 다수의 업계에서 음성표준인터페이스 마크업 언어인 VoiceXML의 적용을 고려하고 있다. VoiceXML은 음성 입출력 디바이스를 통해 사람과 컴퓨터가 의사소통하는 언어이다.[1]

현재 모바일 환경에서의 통신 속도 및 사용자 인터페이스의 편리성을 요구하는 많은 응용 어플리케이션은 음성인식 기능과의 통합 및 연동을 시도하고 있지만 음성인식 시스템은 사용자가 점진적으로 늘어날 때, 빠른 응답시간을 얻기가 힘들어 지거나 음성인식기에 빈번한 장애가 발생하게 된다. 따라서 본 논문에서는 모바일 환경을 고려하여 작은 데이터 전송량 및 서버의 빠른 응답속도를 요구하는 음성인식기를 설계하고 설계된 시스템에 다양한 어플리케이션이 적용되도록 VoiceXML로 구현하였다.

Ⅱ. VoiceXML 해석기 구현

1. VoiceXML 해석기 시스템 구성

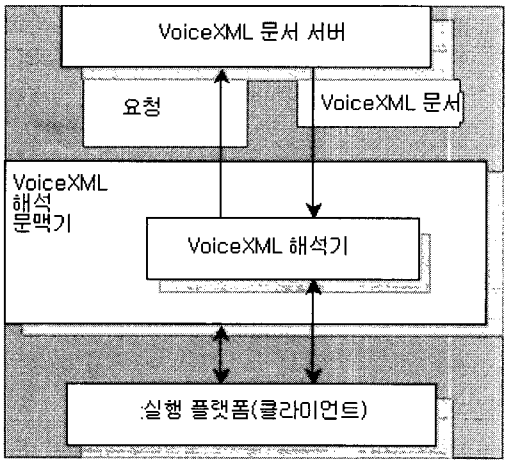
VoiceXML 해석기의 동작을 수행하기 위한 가장 기본적인 시스템은 VoiceXML브라우저, VoiceXML 게이트웨이, VoiceXML 문서서버로 구성되는데 이 중에서 VoiceXML 해석기가 VoiceXML 문서를 해석하는 가장 기본적이고 주된 기능으로서 VoiceXML 게이트웨이에 위치한다. 그리고, 서비스를 위한 VoiceXML 시나리오의 저장소인 문서서버는 웹서버에 위치한다.[2] 〈그림 2.1〉은 이러한 구조를 도식화한 그림이다.



〈그림 2.1〉 VoiceXML 해석기 시스템

2. VoiceXML 해석기 컨텍스트

VoiceXML 해석기는 사용자 음성입력의 시작과 끝을 찾아내며, VoiceXML 문서를 가져와서, 문서의 특정 시나리오에 맞게 자체 로직에 의해 음성인식 엔진과 텍스트 음성출력 엔진에 접목시켜 사용자와의 대화를 제어하는 역할을 수행한다. 그림 2.2는 VoiceXML해석기 컨텍스트를 나타내는 것이다.



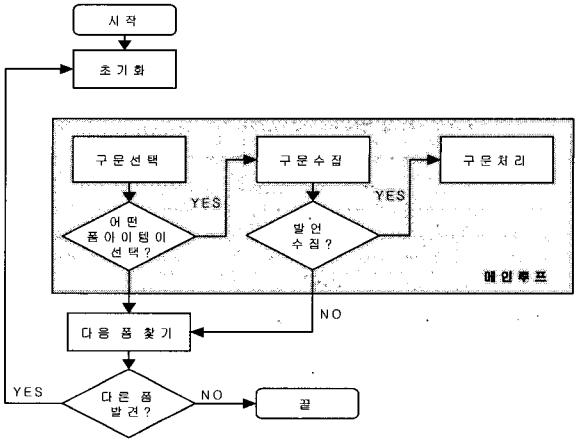
〈그림 2.2〉 VoiceXML의 해석기의 구조

VoiceXML 해석기 컨텍스트는 사용자의 특별한 어문을 뽑아내기 위해서 항상 대기하고 있다. 실행플랫폼은 VoiceXML 해석 컨텍스트와 해석기에 의해 제어되는데, VoiceXML 해석기 컨텍스트는 들어오는 요청을 찾아내는 역할을 하고, 초기화된 VoiceXML 문서를 취하여 요청에 답을 한다. VoiceXML 해석기는 대답 다음의 대화를 생산해 낸다. 문서들의 집합인 어플리케이션 또는 VoiceXML 문서는 회화에서의 명백한 상태를 구성한다.

3. VoiceXML 어플리케이션과 문서

VoiceXML 문서들의 집합인 어플리케이션 또는 VoiceXML 문서는 회화에서의 명백한 상태를 구성한다.[3]

VoiceXML 어플리케이션은 여러개의 문서가 모여 이루어지게 된다. 이는 루트 문서와 어플리케이션의 속성을 통해 루트와 연계되는 다른 문서로 구성된다. VoiceXML의 해석 알고리즘 FIA(Form Interpreter Algorithm)은 사용자와 VoiceXML 폼 또는 메뉴간의 상호작용을 끌어낸다. 이는 FIA를 사용하여 문서에 있는 폼과 폼 아이템의 실행순서를 결정하게 된다. FIA의 다이어그램은 그림 2.3과 같다.

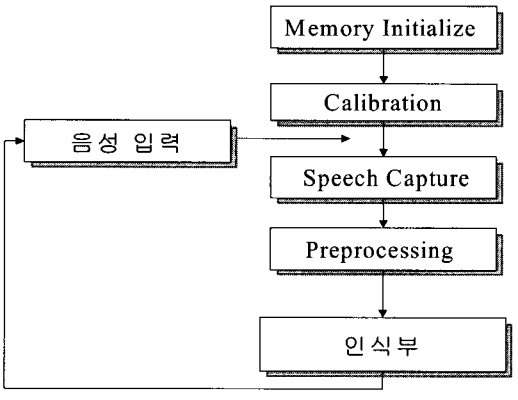


〈그림 2.152〉 FIA 다이어그램

III. 음성인식기 구현

1. 전처리

대부분의 음성인식 구조는 음성을 분석하며 음성신호로부터 음성특징 벡터를 뽑고 나서 통계적인 패턴 분류를 수행한다. 음성분석 알고리즘은 음성신호로부터 프레임 특징벡터를 시간 순으로 뽑아낸다. 통계적 분류 방법은 여러면에서 좋은 장점을 가지고 있다. 음성인식 알고리즘이 최적화되어 입력 프레임에 최적의 PLU 열을 찾아낼 수 있다는 것과 학습 데이터베이스로부터 자동적으로 PLU 모델을 학습하는 과정이 존재한다. 음성 전처리 및 특징추출은 인식기가 사용될 메모리 초기화, 현재 음성 입/출력 장치의 배경잡음 조절, 음성구간검출, 전처리의 과정으로 이뤄진다.[4]

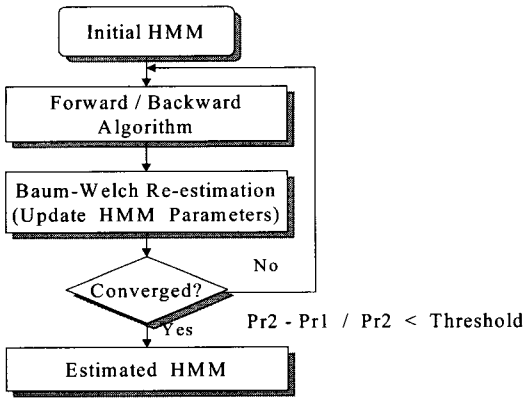


〈그림 3.1〉 전처리의 과정

2. HTK를 이용한 학습과 인식

학습은 녹음된 음성 데이터를 이용하여 HTK를 통해 훈련 DB를 구축하였으며, 학습 과정은 전-후향 과

바움웰츠 알고리듬을 이용해서 음소를 모델링하고, 각 음소 상태에 대한 평균과 분산 그리고 상태 천이 확률 등을 추출하게 되는데 그림 3.2는 학습과정의 흐름도를 나타낸 것이다.[4]



〈그림 3.2〉 HMM의 학습과정 흐름도

인식은 서브워드 단위로 이루어져 있는데 서브워드는 한 어절을 음소단위로 구분하여 놓은 것으로 음성에서는 한 개의 음소만으로 이루는 모노폰과 오른쪽이나 왼쪽에 어떤 음소가 오느냐에 따른 두 개 음소 단위의 바이폰, 그리고 초성, 중성, 종성의 트라이폰 구조로 나누어 구성하고 있다.

Ⅳ. 실험 및 결과

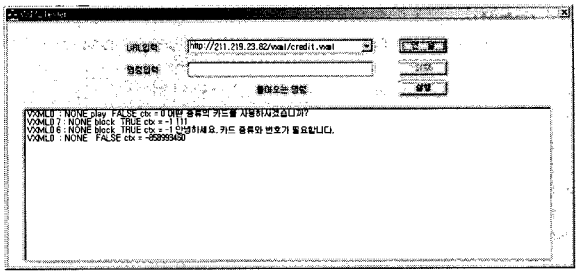
1. VoiceXML 해석기 H/W 사양 및 실험결과

VoiceXML 해석기 시스템은 VoiceXML 브라우저, 게이트웨이, 문서서버로 이루어져 있다. 클라이언트에는 VoiceXML 브라우저를 설치하고, 서버에 VoiceXML 게이트웨이와 문서서버를 함께 설치하였다. 클라이언트와 서버의 하드웨어 사양은 〈표 4.1〉에 정리하였다.

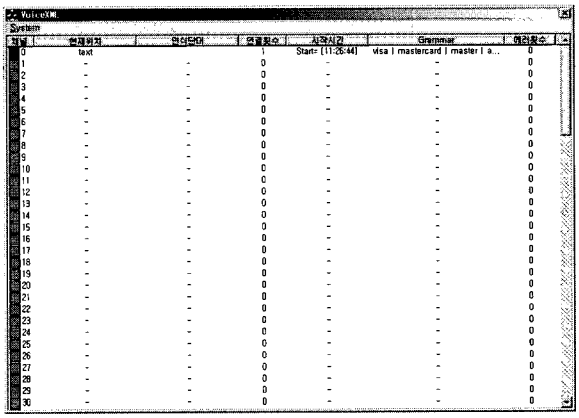
〈표 4.1〉 VoiceXML 해석기 하드웨어 사양

클라이언트 컴퓨터	
Processor	인텔 펜티엄 II
Main Memory	128MB
Network	115.2 kbps
서버 컴퓨터	
Processor	Intel Pentinum III 733MHz
Main Memory	256MB
Network	10Mbps LAN

VoiceXML 해석기의 동작을 실험하기 위해서는 먼저 VoiceXML 게이트웨이를 실행한다. VoiceXML 게이트웨이는 총 120개의 동시연결이 가능하도록 되어 있으며, 각 클라이언트와 연결되면, 채널이 할당되어 연결상태를 모니터링 할 수 있다. 〈그림 4.1〉은 VoiceXML 브라우저의 동작화면을 나타내고 있으며, 〈그림 4.2〉는 VoiceXML 브라우저의 요청에 따른 게이트웨이의 동작 화면이다.



〈그림 4.1〉 VoiceXML 브라우저 동작화면



〈그림 4.2〉 VoiceXML 게이트웨이 동작화면

실험 시나리오를 정리하면 다음과 같다.

① 사용자는 VoiceXML 브라우저를 통해 VoiceXML 게이트웨이로 접속을 하여, 원하는 VoiceXML 문서가 있는 URL을 보낸다.
URL입력 부분에 URL을 입력하고 "연결" 버튼을 클릭한다. (그림 4.1)

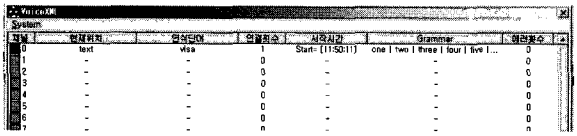
② 게이트웨이는 클라이언트와의 연결요청을 통해 연결이 설정되면 채널을 할당한다. (그림 4.2)

③ 게이트웨이는 URL의 VoiceXML 문서를 가져와 문법을 점검한다. 시나리오에 따라 VoiceXML 브라우저로 질의를 한다. (그림 4.2)

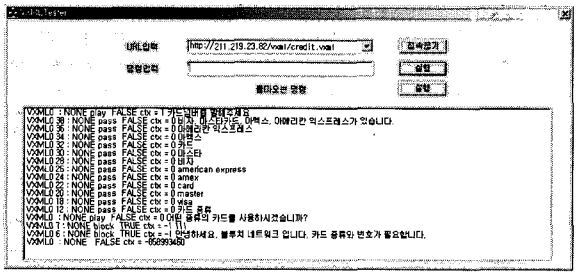
④ 브라우저는 게이트웨이의 질의를 화면으로 출력하고 사용자의 입력을 기다린다. 화면에 "어떤 종류의 카드번호를 사용하시겠습니까?"라는 질의가 출력된다.(그림 4.1)

⑤ 브라우저에서 사용자는 질의에 대한 응답을 명령입력 박스에 입력하여 실행버튼을 클릭한다. "visa"를 입력하고 실행버튼을 클릭하면 게이트웨이에서는

"visa"라는 입력값을 인식한다.(그림 4.3) 그리고, 그에 해당하는 다음 응답을 클라이언트에게 보내준다. VoiceXML 브라우저 화면에 "카드넘버를 말해주세요"라는 질의가 출력된다. (그림 4.4)



〈그림 4.3〉 VoiceXML 게이트웨이의 동작화면 2



〈그림 4.4〉 VoiceXML 브라우저 동작화면 2

이와 같이 시나리오에 따라 VoiceXML 브라우저와 VoiceXML 게이트웨이 사이의 대화가 형성된다. 본 실험의 VoiceXML 해석기는 게이트웨이에 음성 입출력 부분이 포함되지 않은 형태이기 때문에 브라우저 역시 문자위주의 인터페이스로 구성되어 있다.

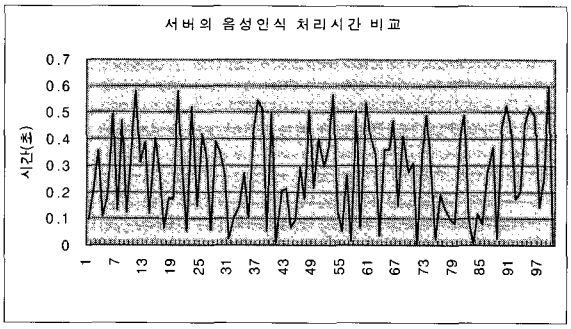
2. 음성인식기 H/W 사양 및 실험결과

모바일 환경을 구현하기 위해 노트북에 핸드폰을 연결하여 실험을 하였다. 노트북은 삼성 sens640을 사용하였으며, 핸드폰은 삼성 SCH 350을 사용하였다. 그리고 마이크를 노트북에 연결하여 음성인식에 이용하였다. 또한 서버에서 음성을 위한 Audio 클래스는 SoundBite 1.0을 사용하였다. 클라이언트와 서버의 하드웨어 사양은 표 4.2에 정리하였다.

〈표 4.2〉 음성인식기 하드웨어 사양

클라이언트 컴퓨터	
Processor	인텔 펜티엄 II
Main Memory	128MB
Sound	16Bit Support
Microphone	Condenser MIC 200 Ω
Network	115.2 kbps
서버 컴퓨터	
Processor	Intel Pentinum III 733MHz
Main Memory	256MB
Network	10Mbps LAN

음성 DB구축은 52명의 남성화자가 조용한 연구실 환경에서 2회 발성한 10개 단어를 녹음하였으며, 인식기에 사용될 수 있도록 음성은 구간 검출 알고리즘을 이용하여 시작점과 끝점을 검출하여 동기화를 이루였고, 8KHz, 16Bit 양자화를 거쳐 Mono형태의 wave PCM 파일로 저장하였다. 사용자의 단어입력에 대한 길이는 0.4-1.7sec초이고, 초당 100번 샘플링을 하므로 초당 40-170 프레임을 사용하였다. 실험에서 약 64KB 정도의 데이터가 서버에 전송되었으며, 200회 실험 시의 서버의 음성인식 처리시간은 평균 0.3초의 처리시간을 소요하였다. 그림 4.5는 서버의 음성 처리시간을 비교하였다.



〈그림 4.5〉 서버의 음성 처리시간 비교

V. 결 론

본 논문에서 제시한 VoiceXML을 이용한 서버측 음성인식 기술은 모바일환경에 적응하도록 서버/클라이언트 모델로 구현되었다. 추후 개발된 VoiceXML 해석기와 서버측 음성인식기를 연계시켜서 VoiceXML 게이트웨이를 구현한 뒤 VoiceXML 브라우저에서 음성 입출력 부분을 추가하여 음성인터페이스를 구현할 것이다.

참고문헌

1. Voice eXtensible Markup Language (VoiceXML TM) version 1.0, W3C 2000.5.3
2. 강유 · 김동준, "early adopter VoiceXML", wrox, 2002.2.
3. 박섭형, "음성 웹 어플리케이션 구축을 위한 VoiceXML", 한빛미디어, 2001.12.
4. 이재왕, "인터넷상에서의 한국어 음성인식", 석사논문, 동신대학교, 1999