모바일 명함 검색을 위한 음성인식시스템 구현

홍인숙*, 고유정*, 김윤중*
*한밭대학교 컴퓨터 공학과
e-mail : ishong@hanbat.ac.kr

A Development of Speech Recognition System for Mobile Card Search

In-Suk Hong*, You-Jung Ko*, Yoon-Joong Kim*
*Dept of Computer Engineering, Hanbat University

요 약

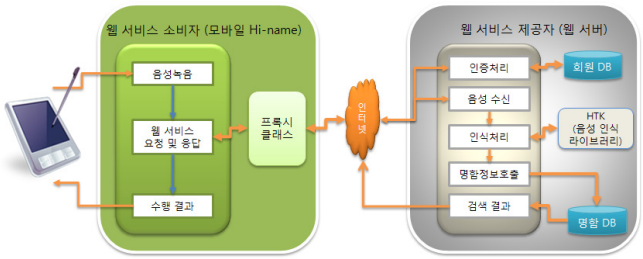
모바일 명함 관리 시스템은 간편하게 모바일 기기를 이용하여 명함을 등록하고 검색할 수 있으나 모바일 기기의 특징상 화면이 작고 정보를 이용하기 위해서는 펜을 이용하여 검색어를 입력해야하는 불편함이 있다. 이를 해결하기 위해 명령을 음성으로 처리하고자하는 VUI(Voice User Interface)의 필요성이 증가하였다. 또한 모바일 기기의 메모리 공간상의 제약으로 인한 음성인식엔진 탑재의 어려움이 있다. 이에 본 논문에서는 모바일 단말기로부터 음성을 입력받아 인식결과를 모바일 단말기로 되돌려 주는 음성인식 시스템을 구축하고 본 인식시스템과 모바일 클라이언트 시스템을 분산처리 가능한 웹서비스 환경으로 구성하였다.

1. 서론

모바일 환경에서의 개인용 모바일 명함관리시스템[1]은 모바일 기기를 이용하여 언제, 어디서나 명함을 인식하고 검색할 수 있는 제품이다. 모바일 명함관리시스템은 기존의 오프라인 위주로 수행하던 명함관리를 원격에서 간편하게 관리하고 검색할 수 있지만 모바일 기기의 특징상 화면이 작아 펜이나 스타일러스로 검색어를 클릭하는 불편함이 발생하였다. 이에 본 논문에서는 이동 단말기에서 편리하게 사용될 수 있는 입력 모드의 하나로서 음성을 이용하여 VUI 처리가 가능한 음성인식 시스템을 구현하고자 한다. 또한 작은 모바일 기기에서 처리할 수 있는 웹서비스 환경을 구축하여 음성인식 시스템을 제공하고자 한다[2][3]. 구현된 음성 인식 시스템의 결과를 평가하기 위하여 음성 명령으로 명함을 검색할 수 있는 시스템을 프로토타입으로 개발하였다. 인식율을 평가하기 위해서 SiTEC 에서 제공하는 음성 DB(DataBase)[4] 일부를 이용하여 음향 모델을 생성한 후 교내 무선 인터넷망을 통하여 발성한 음성을 송신하여 실시간 인식하였다. 실험 결과 95%의 인식 결과를 얻을 수 있었다.

2. 명함관리시스템을 위한 음성인식 시스템 구성

본 논문에서 구현한 음성인식 시스템의 전체 구성도는 (그림 1)과 같다. 음성인식 시스템은 웹서비스 제공자인 웹 서버와 웹서비스 소비자인 모바일 기기로 구성되어 있다. 웹 서비스 제공자는 모바일기기에서 녹음된 음성을 수신 받아 인식하고, 인식 결과를 가지고 명함을 검색하여 결과를 반환한다. 웹 서비스 소비자는 모바일기기에서 명령을 녹음하고, 녹음된 명령을 웹서비스 제공자에게 전송한 후 인식결과에 따른 명령을 수행한다.



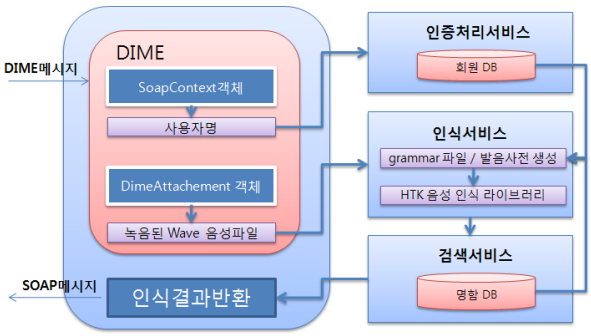
(그림 1) 명함관리시스템을 위한 음성인식 전체시스템 구성도

모바일 명함관리시스템 사용자는 웹서비스 소비자가 탑재된 모바일 기기를 이용하여 사용자정보와 함께 명함정보를 검색하고자하는 이름을 녹음하여 음성인식 웹서비스로 전송한다. 음성 인식 웹 서비스 제공자는 인증처리 후 인증된 사용자를 위한 인식 문법 파일과 발음 사전을 새로이 생성한다. 음성이 들어오면 음성파일을 인식하여 전화번호 등의 검색정보를 모바일 기기로 응답한다.

웹 서비스 소비자는 모바일 기기에 탑재된 명함 관리 프로그램[5]에 HTK를 이용한 음성인식기의 서비스를 호출하는 버튼을 추가하여 구현한 것이다.

3. 웹 서비스 제공자

웹 서비스 제공자는 (그림 2)와 같이 웹 서비스 소비자에게 인증 처리, 인식 처리, 검색 서비스를 제공한다.



(그림 2) 웹 서비스 제공자

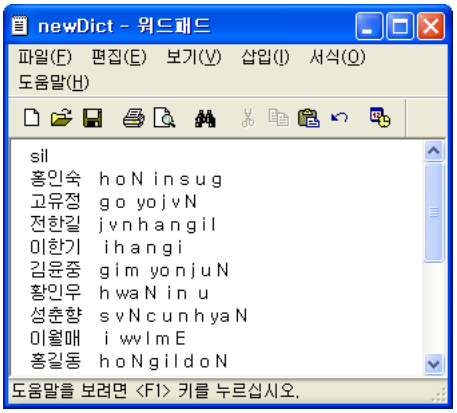
먼저 웹 서비스 제공자는 웹 서비스 소비자로부터 DIME (Direct Internet Message Encapsulation)[6] 메시지를 전달 받아 사용자 정보와 음성 파일로 분리한다. 사용자 정보는 인증처리 서비스를 이용하여 음성 인식 웹 서비스 권한을 부여 받는다. 인증처리가 완료되면 사용자의 정보를 가지고 명함 DB에서 인식대상들을 처리한 다음 추출된 단어 리스트를 이용하여 인식을 위한 문법 파일과 발음사전을 생성한다. 문법 파일은 EBNF 형식으로 구성되고, 발음 사전은 자소분리 테이블을 이용하여 인식대상 어휘와 인색 대상 어휘를 구성하는 음소 모델 열의 쌍으로 생성된다. 이 과정은 사용자가 처음 웹 서비스를 요청할 때 한 번 수행된다.

음성파일은 인식 서비스의 입력으로 사용되며 인식한 결과에 따라 검색 서비스의 명함 DB에서 일치하는 정보를 찾아 웹 서비스 소비자에게 응답한다.

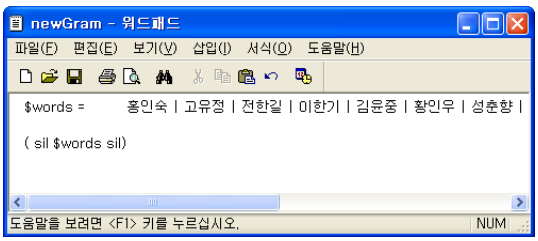
4. 음성 인식기

인식 단위 훈련 및 인식은 HTK 3.2(HMM Tool Kit)[7][8][9]를 이용하여 구현하였다. HTK는 HMM을 기반으로 한 음성 인식 시스템 구현 및 실험을 위한 상용도구이다. 총 452개의 단어로 이루어진 마이크 녹음의 남성 화자에 대해서만 훈련을 하였으며, 음성 신호를 표현하기 위해 음성 데이터에 25ms 길이의 해밍 윈도우를 씌우고 10ms마다 12차 MFCC와 에너지 및 그에 해당하는 델타 에너지 등의 특징 파라미터를 구하여 구성한 39차 특징 벡터를 사용하였고, 음성의 특징 파라미터 추출은 HCopy 도구에 의해 수행하였다. 음향 모델 생성을 위해 3 state left-to -right 모델을 사용하였다. HTK Tool Kit을 이용해 음향 및 언어모델을 훈련한 뒤 기본 디코더인 HVite 디코더를 이용해 음성 인식기를 구현하였다. 훈련과정을 통해 얻은 각 음소 상태에 대한 평균과 분산 값, 상태 천이 확률을 이용하여 인식 모델과 비교 인식결과를 출력하였다. 훈련에 사용된 음소의 수는 43개(묵음 제외)이다.[10]

명함 DB의 내용은 수시로 변경되어 질 수 있으므로, 단어 단위가 아닌 음소 단위로 음성을 모델링하여 음성을 인식하고자 하였다. 음소 단위로 모델링한 결과 단어 내 혹은 단어 간의 조음현상을 고려하지 않아 인식율의 저하를 가져오는 현상을 볼 수 있었다. 이에 저하된 인식률을 높이기 위해 명함DB에 등록된 모든 단어를 인식대상으로 구성하지 않고 인증된 사용자에 관한 정보에 대해서만 그 인식대상을 제한하였다. 사용자가 변경되면 새롭게 인식을 위한 문법파일(Grammar)과 발음사전(newDict)이 생성된다. Grammar 파일은 그림과 같은 형태로 구성되었으며, 표준 EBNF에 따른다.



(그림 3) 새로 만들어진 발음사전(newDict)



(그림 4) 새로 만들어진 문법파일(Grammar)

또한 본 논문에서 구현된 음성 인식기의 비교 실험을 위해 정보기술업체인 에스엘투(SL2)에서 구축한 화자독립 방식의 가변어휘인식을 지원하는 VoiceLink 3.4를 이용하였다. [11]

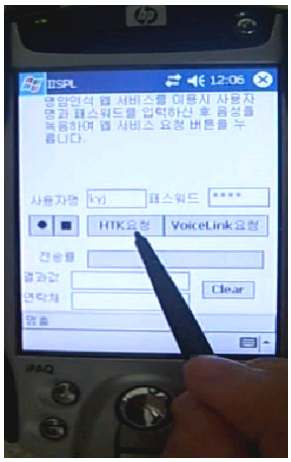
5. 실험 및 결과

본 시스템의 성능을 알아보기 위해 실험은 Open Test로 <표 1>과 같은 환경에서 수행하였으며 실험에 이용한 기종은 HP iPAQ H5420이다.

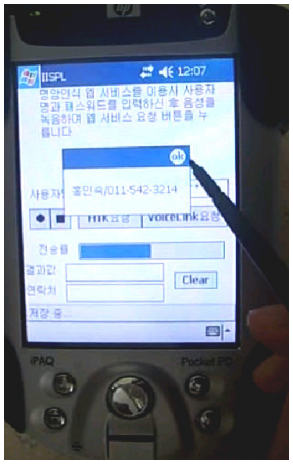
<표 1> 테스트 환경

데이터		음성파일	녹음장소
훈련용 데이터		16bit , 채널1(mono) 16000kHz, NIST 방식	방음시설
평가용데이터	테스트1 (그룹1)	16bit , 채널1(mono) 16000kHz, NIST 방식	방음시설
	테스트2 (그룹2)	16bit , 채널1(mono) 16000kHz, PCM 방식	일반사무실
	테스트3 (그룹2)	16bit , 채널1(mono) 16000kHz, PCM 방식	일반강의실

(그림 5)은 모바일 명함 관리 시스템의 음성 인식을 위한 인터페이스 화면이며, (그림 6)은 음성인식시스템의 결과를 나타내는 화면이다.



(그림 5) 인터페이스화면



(그림 6) 음성인식 결과

음성 DB[4]로 사용된 훈련용, 평가용 데이터는 총 38명의 남성(그룹1) 음성 데이터로 1인당 452개의 단어를 1회 발성하였다. Open Test로 30명은 훈련용으로, 나머지 8명은 평가용으로 사용하였다. 두 개의 인식기의 성능을 비교하기 위해 음성DB에 사용된 발성자를 제외한 새로운 발성자, 성인 남자 2명, 성인 여자3명을 선정(그룹2)하였다. 훈련에 참가하지 않은 사용자(그룹1) 452(단어/명) * 8(명) = 3616개 단어를 Test1 , 훈련에 참가하지 않은 사용자(그룹2) 20(단어/명) * 5(명) * 5(회) = 500 개 단어를 Test2 , Test2와 동일한 화자(그룹2) 10(단어/명) * 5(명) * 2(회) 개 단어를 Test3으로 사용하여, Test1은 구현된 음성인식기의 성능을, Test2는 유선 상에서의 음성인식기의 성능을, Test3은 무선 상에서의 음성인식기의 성능을 비교 실험하였다.

성능 비교 결과는 다음 <표 2>과 같다.

<표 2> 인식기에 대한 인식률 비교

인식기	데이터	인식환경	인식수/ 실험횟수	인식률
HTK	훈련용	유선	10296/10848	94.9%
	테스트1	유선	3064/3616	84.7%
	테스트2	유선	480/500	96%
	테스트3	무선	94/100	94%
VoiceLink	테스트2	유선	482/500	96.4%
	테스트3	무선	96/100	96%

인식 대상을 제한한 결과 그 인식률이 향상되었음을 알 수 있었다.

6. 결론

본 논문에서는 명령 입력이 불편한 모바일 기기에 음성인식 기술을 이용한 VUI를 구현함으로써 사용자와의 인터페이스를 편리하게 제공하였다. 또한 음성인식 엔진을 모바일 기기에 탑재하지 아니함으로써 인식 시 요구되는 자원 이용의 제약성을 극복하였다[2][3]. 모바일 명함관리 시스템을 위한 음성인식 웹 서비스 환경을 구축함에 따라 웹서비스 소비자가 탑재된 어떤 모바일 기기에서도 언제, 어디서나 인식 시스템을 사용할 수 있으며, 인식 대상을 웹 서비스 제공자에서 수정하기 때문에 모든 모바일 기기에 최신의 정보로 적용되어 질 수 있다. 또한 그 인식 대상을 제한하였기 때문에 잡음이 많은 무선 환경에서의 인식률도 높일 수 있었다.

추후 인식기의 성능을 높이기 위해 음소 모델을 음소단위가 아닌 음소의 좌우 음운현상을 고려한 문맥 종속형 음소 모델(tri-phone 단위)을 이용하는 인식기를 구현하고자 한다. 단, 음향학적 모델 훈련 과정에서 나타나지 않은 음소 모델이 인식대상어휘에서 나타나는 unseen model 문제가 발생할 수 있는 문제점에 대한 해결책도 제시하고자 한다. 또한 사용자 인증 처리를 위해 사용자 정보를 펜으로 입력해야하는 불편함을 없애고자 화자인식 기능을 추가하여 사용자에게 좀 더 편리한 시스템을 구현하고자 한다.

참고문헌

[1] Hi-Name, http://www.hiname.net/Roger S. Pressman "Software Engineering A Practltiners' Approach" 3rd Ed. McGraw Hill

[2] 임수호 외 5명, "무선 네트워크 환경 하에서의 음성인식에 관한 고찰" , 한국음향학회 학술발표대회 논문집 제23권 제2호, 2004.

[3] 박영주 외 5인, "모바일환경에서의 VoiceXML을 이용한 cs 모델에 관한연구", 음성통신 및 신호처리 학술대회 논문집 19권 1호, 2002.

[4] 김봉완, 김종진, 김선태, 이용주, “공동 이용을 위한 음성 DB의 설계 및 구축에 관한 연구”, 한국 음향학회지 제16권, 제4호, pp35-41, 1997 .
[5] 고유정, 홍인숙, 김윤중, 송은숙, “모바일 하이네임을 위한 음성인식 웹 서비스 환경 구축”, 한국정보처리학회 추계학술발표대회 논문집 제15권 제2호, 2008. 11
[6] Jeannine Hall Gailey, “Sending Files, Attachments, and SOAP Messages Via Direct Internet Message Encapsulation”,http://msdn.microsoft.com/en-us/magazine/cc188797.aspx
[7] HMM Tool Kit , http://htk.eng.cam.ac.uk/
[8] Steve Uoung, "The HTK Book (for HTK Version 3.2)
[9] Steve Young, "The General Use of tying in Phineme-Based HMM Speech Recognisers" , IEEE International Conference on Acoustics, Speech and Signal processing , pp.569-572 , 1992.
[10] 홍인숙, “HTK를 이용한 음성 인식 시스템 구현“ , "http://home.wins.or.kr", 2008. 02
[11] VoiceLink, http://www.vlink.co.kr/