휴대폰용 멀티모달 인터페이스 개발

키패드, 모션, 음성인식을 결합한 멀티모달 인터페이스

Development of a multimodal interface for mobile phones

김원우, Wonwoo Kim

요약 휴대폰은 현대 생활에 없어서는 안 될 개인화 단말기가 되었으며, 그 위에서 다양한 디바이스, 컨텐츠 및 서비스의 컨버전스가 이루어지고 있다. 그러한 다양하고 복잡한 기능과 대용량 컨텐츠 및 정보를 효과적으로 검색하고 사용할 수 있는 수단에 대한 연구도 활발히 진행되고 있다. 본 연구는 휴대폰 상에서 음성, 키패드, 모션을 이용하여 한글 단어를 입력하는 새로운 인터페이스를 개발하고, 이를 응용한 전화걸기 애플리케이션을 통하여 그 사용성과 효과를 검증하는 것을 목적으로 한다. 개발된 멀티모달 인터페이스는 복잡한 메뉴 트리와 깊이를 한 번에 접근할 수 있는 음성 인터페이스의 장점을 수용하면서 인식률 및 인식시간을 개선하였다.

Abstract The purpose of this paper is to introduce a multimodal interface for mobile phones and to verify its feasibility. The multimodal interface integrates multiple input devices together including speech, keypad and motion, It can enhance the rate and time for speech recognition, and shorten the menu depth.

해심어: Multimodal, Interface, Speech, Motion

1. 서론

국내 휴대폰 가입자가 4천만 명을 넘어서는 시대에 이르러 휴대폰 기능은 단순한 전화 연결에서부터 카메라, 인터넷, MP3, 게임 등과 계속 컨버전스되고 있으며 데이터나 컨텐츠를 저장하기 위한 기억 용량도 수 기가에 이르렀다.

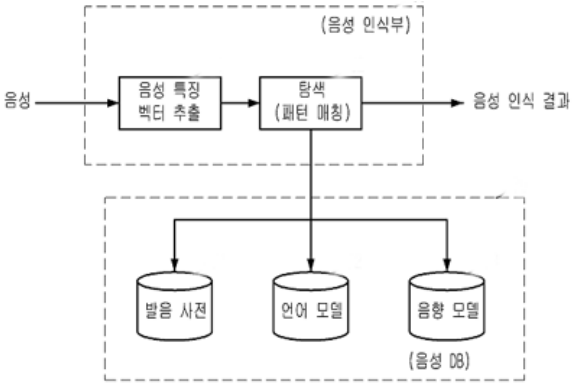
이에 따라 다양하고 복잡한 기능과 대용량 컨텐츠 및 정보를 쉽게 검색하고 사용할 수 있는 인터페이스에 대한 연구가 활발히 진행되고 있다.

본 논문은 그러한 연구 중의 하나로서, 휴대폰 상에서 음성, 키패드 및 모션을 이용하여 한글 단어를 입력하는 새로운 멀티모달 인터페이스를 제시하고, 이를 응용한 전화걸기 애플리케이션을 통하여 그 사용성과 효과를 살펴보았다.

2. 음성인식과 멀티모달 인터페이스

2.1 음성인식 인터페이스

사람의 음성은 가장 자연스러운 의사소통 수단으로써, 오래 전부터 영화 스타워즈에서와 같이 기계와의 자연스러운 대화를 꿈꾸어 왔다. 최근 MS의 빌 게이츠가 음성인식 기술에 눈독을 들이고 있다는 기사는 더욱 그 가치가 높아졌음을 시사하고 있다.



[그림 1] 일반적인 음성인식 처리과정

음성인식은 그림 1에서 보는 바와 같이 마이크를 통하여 들어온 사용자 음성을 HMM(Hidden Markov Model)을 이

용하며 특징 벡터를 추출하고, 추출된 음성 특징 벡터에 대해 음성DB 내 발음사전, 언어모델 및 음향모델을 탐색해 패턴매칭을 수행함으로써 입력된 음성에 대응되는 단어를 찾는 과정을 수행한다.

이러한 음성인식 기술은 주변 잡음이 적은 실험실에서는 90% 이상의 좋은 성능을 보이고 있지만, 잡음이 심한 복도, 전시장, 회의장, 차량 등에서는 인식률이 현저히 떨어지는 한계를 나타내고 있어서 실용화 및 대중화의 발목을 잡고 있으며, 이를 보완하기 위한 다양한 시도가 이루어지고 있다.

또한 마이크를 통해 들어온 소리 중에서 어떤 것이 발화자의 음성인가, 그리고 해당 음성의 어느 부분에서부터 실제 의미가 있는 음성인가를 찾아내는 문제에 대해서도 아직 완전한 솔루션은 없다고 하겠다.

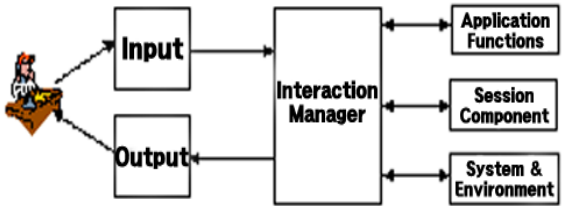
특히 한국어는 다른 여러 나라의 언어에 비해 상대적으로 낮은 인식률을 보이고 있는데, 예를 들어, 영어의 경우 음성인식의 기초가 되는 '단어'로 문장이 구성되어 있는 반면, 한국어는 한 글자로 된 조사 등으로 구조를 좀 더 복잡하게 만들기 때문인 것으로 이해된다. 또한 신호 에너지가 낮아 인식이 어려운 파찰음 문제도 풀어야 할 숙제라고 할 수 있다.

2.2 멀티모달 인터페이스

멀티모달 인터페이스는 인간과 기계(컴퓨터, 단말기 등)의 통신을 위하여 음성, 키보드, 펜, 터치, 센서 등 다양한 모드를 함께 사용하는 것을 말한다.

멀티모달 인터페이스는 유비쿼터스 환경에서 사용자를 인지하는 기술로 많이 연구되고 있으며, 키보드, 터치패드와 같은 기존 인터페이스에 음성인식 기술을 접목하는 연구개발도 많이 이루어지고 있다.

모달리티가 추가됨에 따라 사용자 인터페이스의 구현은 더욱 복잡하고 까다로워지는데, W3C(World Wide Web Consortium)에서는 개발자들이 쉽게 멀티모달 서비스 및 단말을 개발할 수 있도록 멀티모달 인터렉션 프레임워크 표준안을 개발하고 있다. 그림 2은 W3C에서 제안하고 있는 멀티모달 프레임워크를 보여준다.



[그림 2] W3C의 멀티모달 프레임워크

3. 휴대폰용 멀티모달 인터페이스 개발

3.1 휴대폰용 멀티모달 인터페이스

본 논문에서 제안한 휴대폰용 멀티모달 인터페이스는 음성인식을 기반으로 하여 키패드와 모션을 결합하였다. 먼저 동작인식을 통하여 미리 정의된 애플리케이션을 수행하고, 키패드와 음성인식을 통하여 해당 애플리케이션에 필요한 패러미터를 전달하는 방식이다.

이해를 쉽게 하기 위하여 전화번호부 검색을 통한 전화걸기 애플리케이션을 예로 들어 설명한다. 먼저 휴대폰을 앞뒤로 흔들면 애플리케이션이 구동되어 그림 3과 같은 화면이 나타난다.

[그림 3] 애플리케이션 초기 구동화면

초기 화면에서 전화하고자 하는 착신자의 이름을 발화하거나, 인식률을 높이기 위해 해당 이름의 첫 번째 초성 자음을 키패드 상에서 누른 후, 이름을 발화한다. 예를 들어, 홍길동을 입력하고자 하는 경우, 'ㅎ'을 누르고 '홍길동'을 말하면 된다.

음성인식이 끝나면 그림 4와 같은 화면이 출력된다.

[그림 4] 인식결과 화면

음성인식 결과는 정확도에 따라 1명만 출력하거나, 정확도 순서에 따라 3명을 출력한다. 음성인식 결과의 정확도가

높아 1명만 출력하는 경우에는 바로 해당하는 사람의 전화 번호로 연결을 시도한다.

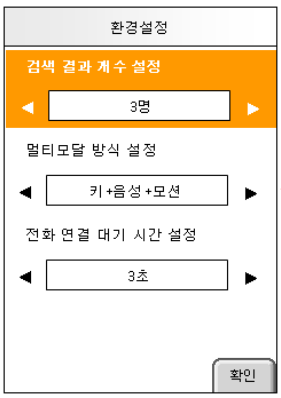
음성인식 결과로 3명의 대상이 출력되는 경우에는 3명 중에서 한 명을 선택하는 과정을 거쳐야 하는데, 여기서는 동작인식 기능을 포함시켰다. 그림 5는 이 과정을 보여준다.

[그림 5] 인식결과 선택과정

음성인식의 결과로 아름, 아람, 아빠가 출력되었고 그 중의 하나를 선택하기 위해 앞 뒤 모션을 사용할 수 있고, 해당하는 키패드 버튼(1~3)을 누를 수도 있으며, '1', '2', '3'을 말할 수도 있다. 터치패드가 가능한 단말기에서는 해당 메뉴 또는 버튼을 클릭하는 것도 포함될 수 있다. 이러한 선택은 곧 바로 전화걸기로 이어진다.

전화걸기는 특정 인식결과가 포커싱된 상태에서 타임아웃에 의해 3초나 5초 후 자동으로 해당 착신자에게 전화를 걸게 되며, 원하는 착신자가 포커싱된 상태에서 좌-우로 흔드는 모션으로 즉시 전화걸기를 수행할 수도 있다.

그림 6은 이 애플리케이션에 적용된 옵션을 보여주고 있는데, 음성인식 결과값을 몇 명을 할 것인가, 사용할 모달리티가 어떤 형태인가, 그리고 음성인식을 완료하고 전화연결까지의 대기시간이 몇 초인가를 설정할 수 있다.



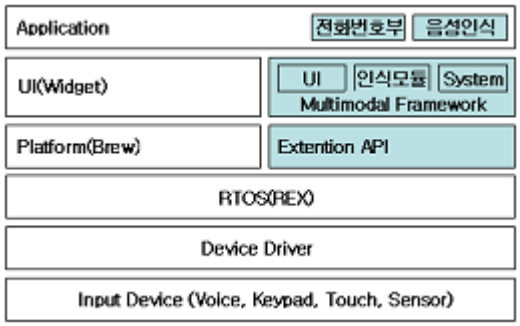
[그림 6] 환경설정 화면

3.2 휴대폰용 멀티모달 인터페이스 개발

멀티모달 인터페이스의 개발환경은 다음과 같다.

- 개발 언어 : C / VC++
- 개발 도구 : Cygwin, VisualStudio
- 개발 플랫폼 : BREW, PocketPC 2003
- 개발 단말기: SPH-B3200, HP IPAQ HX4700

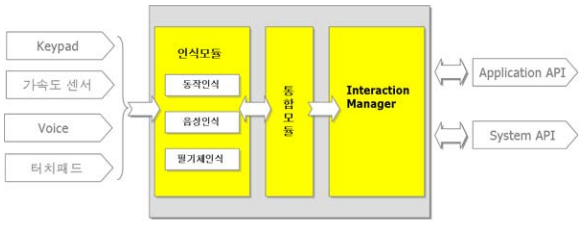
멀티모달 시스템은 그림 7과 같이 음성, 키패드, 터치 및 센서와 같은 입력장치와 이를 인식하는 인식모듈, 그리고 플랫폼 및 애플리케이션으로 구성된다.



[그림 7] 시스템 구성도

색칠한 부분이 본 논문에서 설명하고 있는 멀티모달 인터페이스의 구현범위이다.

멀티모달 플랫폼은 블루(Brew) 기반으로 개발하였으며, 표준화된 UI 및 시스템 연동을 위하여 W3C에서 제시하고 있는 멀티모달 프레임워크를 탑재하였다. 프레임워크의 구성도는 그림 8과 같다.



[그림 8] 멀티모달 프레임워크 구성도

4. 장점 및 효과

개발된 멀티모달 인터페이스는 일반적인 음성인식의 한계를 보완해 주고, 동작인식과의 결합을 통해 새로운 인터페이스 방식을 제안한다.

4.1 음성인식률 개선

음성인식 성능의 개선은 크게 음성인식률의 향상, 음성인식 시간의 감소 및 시작점 검출 오류의 감소으로 설명할 수 있다.

음성인식률 향상은 인식하고자 하는 단어의 첫 번째 초성 자음으로 시작하는 단어들로 인식범위를 축소하여 인식을

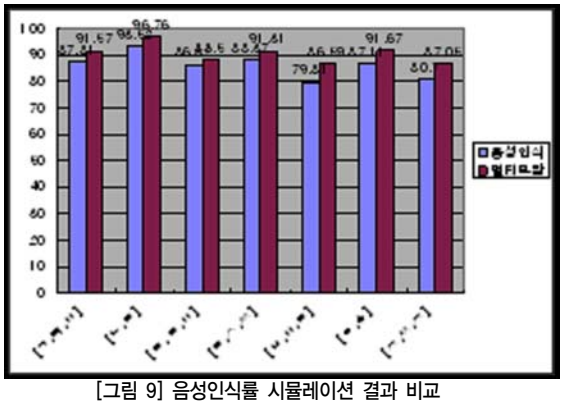
수행하기 때문에 이루어지는데, 이는 음성인식 시간의 감소로도 이어진다.

또한 첫 번째 자음이 입력되는 시점에서 사용자 음성 입력을 기다리기 때문에 음성인식의 어려움 중의 하나인 시작점 검출 측면에서도 도움을 준다.

멀티모달 인터페이스의 음성인식률 향상은 시뮬레이션을 통하여 기 측정한 바 있으며, 약 12,000개의 실제 음성을 사용하여 철도역명을 대상으로 실험한 결과 평균 4.1%의 인식률 향상을 확인하였다.

일반적인 음성인식 방법에 의한 인식률이 85.84%였고, 멀티모달 인터페이스를 적용했을 때의 인식률은 89.94%를 나타내었다.

인식률 개선의 분포에 있어서도 2.18%~6.78%의 비교적 고른 개선이 이루어졌음을 알 수 있었다.(그림 9 참조)



[그림 9] 음성인식률 시뮬레이션 결과 비교

위의 결과는 초성 자음만 알려주고 인식을 수행한 결과이므로, 실제 환경에서는 좀 더 높은 인식률 개선이 가능할 것으로 예상된다.

4.2 음성인식 시간 감소

음성인식률의 향상과 함께 음성인식 측면에서의 또 다른 효과는 음성인식 과정의 대기시간 감소이다. 음성인식 대상 단어의 범위(수)를 축소함에 따른 결과이므로 전체 음성인식 대상의 크기가 클수록 음성인식 시간의 감소도 커짐을 예상할 수 있다.

음성인식 시간의 측정은 휴대폰 상에서 실행하였으며, 정확한 시간 측정을 위하여 프로그래밍을 수행하였다. 음성인식 대상인 DB 인덱스의 크기는 50개, 60개, 70개로 하였으며, 그에 따른 인식시간의 개선을 확인하였다.

일반적인 음성인식 시간 측정은 다음과 같이 진행된다.

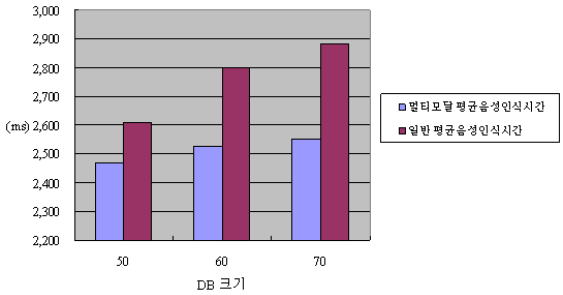
어플리케이션 실행->[전체DB] 선택->원하는 대상을 음성으로 입력->인식결과(대상이름/Fail) 및 시간(ms) 디스플레이

본 논문의 멀티모달 인터페이스에서의 음성인식 시간은 다음과 같이 측정된다.

어플리케이션 실행->[선택DB] 선택->원하는 대상의 자음을 키패드로 입력->해당 대상을 음성으로 입력->인식결과(대상이름/Fail) 및 시간(ms) 디스플레이

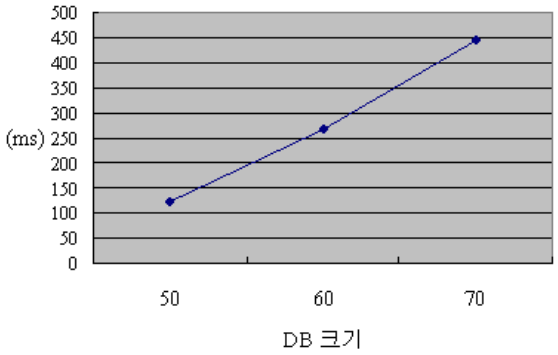
첫 번째 케이스는 마이크를 통하여 입력받은 사용자 음성에 대하여 음성인식 DB 전체에서 일치 또는 유사한 단어를 검색하는 반면, 두 번째 케이스는 먼저 키패드로 입력된 자음을 통해 전체 DB에서 그 입력된 자음으로 시작하는 단어들을 검색한 후, 해당 단어들(범위가 축소된) 내에서 음성인식 결과를 찾는다.

그림 10은 DB크기 별 평균 음성인식 시간을 그래프로 나타낸 것이다. 일반적인 음성인식의 평균 음성인식 시간은 DB인덱스가 50개, 60개, 70개일 때 각각 2,610ms, 2,798ms, 2,882ms를 나타낸 반면, 멀티모달 인터페이스를 적용하였을 경우에는 각각 2,467ms, 2,529ms, 2,552ms를 나타내었다.



[그림 10] 음성인식 DB별 평균 음성인식 시간

이 그래프는 일반적인 음성인식에 비해 멀티모달 인터페이스를 사용하는 것이 음성인식에 걸리는 시간을 크게 단축시킬 수 있음을 보여주고 있으며, 이것은 멀티모달 인터페이스에서의 음성인식의 장점을 더욱 부각시킬 수 있는 효과라고 할 수 있다.



[그림 11] DB크기에 따른 인식시간 개선

또한, 음성인식 시간의 감소 즉, 일반적인 음성인식에 걸린 시간에서 멀티모달 인터페이스를 적용했을 때의 음성인

식 시간의 차를 그래프로 나타내면 그림 11과 같다.

DB 인덱스의 크기가 커짐에 따라, 인식시간 개선이 125ms에서 443ms까지 증가함을 보임으로써, DB 크기가 커질수록 그 효율성이 큰 폭으로 높아질 것임을 알 수 있다.

이러한 점은 휴대폰 내 정보 및 컨텐츠의 용량이 급속히 증가하고 있음을 고려해 볼 때, 크게 주목해야 할 점으로 평가된다.

5. 결론

다양한 휴대용 정보 기기들의 정보 및 콘텐츠를 효과적으로 검색하고 손쉽게 이용하기 위한 수단으로써 음성인식 인터페이스의 중요성이 강조되고 있다. 특히 유비쿼터스 환경의 도래와 함께 비주얼 인터페이스의 제약이 예상되는 바, 그 필요성은 더욱 높아졌다고 할 수 있다.

특히 현재의 음성인식 기술은 주변 잡음에 취약성을 나타내는 등 사용자가 요구하는 성능 및 품질을 만족시키지 못하고 있다.

본 논문은 음성인식에 키보드나 키패드, 동작인식을 접목함으로써 음성인식 단일 기술이 갖는 기술적 한계를 극복하고 미래의 유비쿼터스 환경에 보다 적합한 멀티모달 인터페이스를 설계하고 또한 휴대폰 상에서 구현해 보았으며, 그 효과를 음성인식률, 음성인식 시간, 시작점 검출 및 메뉴트리 간소화 측면에서 살펴보았다.

구현된 멀티모달 인터페이스는 복잡한 다단계 메뉴 트리를 한 번에 접근할 수 있을 뿐만 아니라, 음성 인터페이스의 시작점 힌트와 음성인식의 범위 축소함으로써 전반적인 인식률 향상과 인식시간을 단축하였다.

특히 음성인식의 대상이 될 휴대폰 내 정보 및 컨텐츠의 용량이 급속히 증가하고 있음을 고려해 본다면, 이러한 멀티모달 인터페이스의 개발이 꼭 필요하다고 하겠다.

참고문헌

[1] 김원우, 전호현, “음성/키 패드를 이용한 한글 단어 입력용 멀티모달 인터페이스”, HCI 2007

[2] 구명완, “음성 인터페이스와 멀티모달 인터페이스 “, ITFIND 주간기술동향 통권1193호(2005.4.27)

[3] Ho-Hyun, Jeon, et al., “A Speech Operated Railroad Information & Reservation Service With Multistatage Dialogue”, SST-2000

[4] Multimodal Architecture and Interfaces, W3C Working Draft 11 December 2006, http://www.w3.org/TR/2006/WD-mmi-arch-20061211/