특집

음성 인터페이스 기술의 현황과 전망

김병창(대구가톨릭대학교), 정민우 · 이근배(포항공과대학교)

Ⅰ. 서 론

사람사이의 가장 자연스러운 의사 교환 수단인 음성을 기계와의 인터페이스 수단으로 사용하고자 하는 것은 당연한 일이라 하겠다. 사람은 음성을 사용하여 여러 가지 정보를 주고받게 된다. 1차적으로 사람은 음성을 사용하여 의미를 교환할 수 있으며, 2차적으로는 음성 발화자의 신분, 음성내의 감정 등을 파악할 수 있다. 이러한 능력을 기계에게 부여하고자 하는 것이 음성 인터페이스 기술의 연구 목표라 하겠다. 의미 교환을 위한 음성의 발화는 음성합성기술에 의해 구현되며, 음성 안에 있는 의미의 파악은 음성인식기술에 의해 구현된다. 음성의 발화자를 인지하는 화자 인식기술과 발화자의 감정을 인식하는 감정 인식기술도 연구되고 있다. 또한, 음성합성기술과 음성인식기술은 언어처리기술과 융합되어 좀더 다양하고 편리한 인터페이스를 제공하기도 한다.

1997년 LG전자의 프리웨이라는 브랜드로 김혜수가 "우리집"을 외치며 음성인터페이스 기술의 상용화를 알린 일이 있었다. 이후로 주식시세안내, 철도예약, 일기예보, 음성 다이얼링, 음성에 의한 문서작성 등의 용도로 음성 인터페이스가 활용되기 시작했다. 아직은 음성인식기술에서의 오인식과 음성합성기술에서의 낮은 합성음질은 많은 상품의 상용화를 가로막고 있지만, 현재의 기술수준으로 제한된 영역에서의 실용적인 제품이 꾸준히 출시되고 있다.

다양한 휴대기기의 발달은 다양한 인터페이스 기술에 대한 요구를 증폭시키고 있다. 특히 사람의 눈과 손이 자유스럽지 못한 환경에서는 음성 인터페이스만이 유일한 대안으로 떠오르고 있다.

본 고에서는 음성 인터페이스를 위한 기본 기술인 음성합성기술과 음성인식기술에 대해 설명하고, 각 기술의 음성 인터페이스 활용 현황을 간단하게 살펴본다. 그리고, 음성 인터페이스 기술의 미래에 대해 기술하고 결론을 맺는다.

II. 음성인식 기술

1. 음성인식 기술의 개요

음성인식이란 마이크나 전화로부터 입력받은 음성신호를 단어열로 변환하는 과정이라 말할 수 있다. 인식된 단어는 명령제어, 데이터 입력, 문서 작성 등과 같은 응용의 최종입력으로 사용될 수 있다. 또한, 음성이해와 같은 추가 언어처리를 위한 입력으로도 사용될 수 있다.

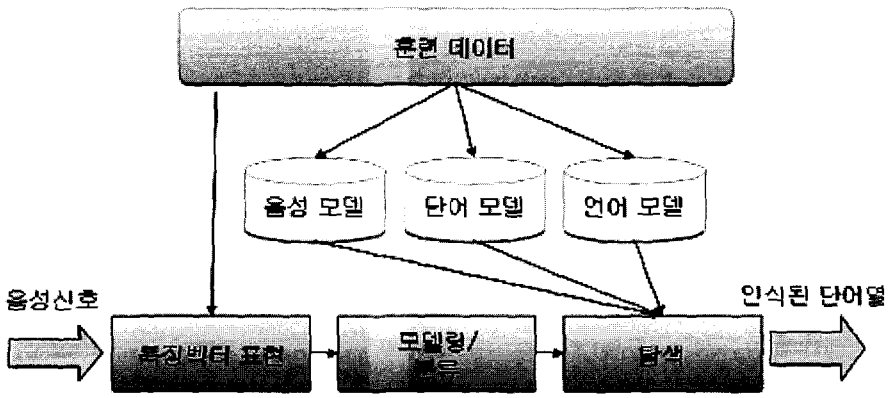
<그림 1>은 음성인식 시스템의 주요 구성을 보여주고 있다. 입력된 음성은 먼저 일정한 시간간격(보통 10-20ms)으로 특징벡터로 표현된다. 이러한 특징벡터는 음성모델, 단어모델, 언어모델 등의 제약조건을 활용한 탐색과정에서 가장 그럴듯한 단어를 탐색하는데 사용된다. 미리 준비된 훈련 데이터는 인식시스템에 사용되는 모델의 값을 결정하는데 사용된다.

이러한 음성인식 시스템의 성능을 비교하는 것은 아주 까다로운 일이며, 보통 다음의 4가지 조건을 고려하여야 한다[2][3].

(1)화자 : 음성인식 대상인 음성의 화자에 대한 조건이다. 화자의 음성이 음성인식 시스템의 음성모델 제작에 사용되었으면 화자 종속이라 한다. 음성모델의 제작에 참여하지 않은 화자의 음성을 인식하는 것을 화자 독립이라 하고, 화자 종속에 비해 화자 독립의 성능이 낮을 수밖에 없다. 하지만, 음성모델의 개발에 사용자가 참여하기는 사실상 불가능하므로, 대부분의 음성인식 시스템은 화자독립이다.

화자 적응은 음성인식 시스템의 음성모델 제작에는 참여하지 않고, 음성인식 시스템의 사용전에 화자의 음성을 사용하여 음성모델을 화자에 적응시키는 방법이다.

(2)발화 스타일 : 하나의 단어를 발화하여 인식하는 고립단어 인식이 있으며, 연속적인 단어의 발화를 인식하는 연속단어 인식이 있다.



〈그림 1〉 음성인식 시스템의 주요 구성[1]

기술적 난이도는 연속단어 인식이 더 어려우며, 연속단어 인식에도 낭독체 인식보다 문법적인 제약이 약한 대화체 인식이 더 어렵다.

(3)응용 도메인 : 단독 숫자를 인식대상으로 하는 음성인식은 가장 쉬운 응용도메인에 속한다. 일반적으로 주식거래나 음성 다이얼링 등 제한된 도메인을 응용도메인으로 정하고 있다. 신문, TV방송, 무제한 도메인 등 아직은 현재의 음성인식 기술을 응용하기에 어려운 응용도메인이 많다. 응용도메인에 따라 입력되는 음성의 문법구조와 인식대상 단어의 크기가 결정된다. 문법구조는 음성인식의 대상에 대한 제약으로 사용하며, 문법적 제약이 강할수록 인식기의 성능이 높아진다. 단어의 크기는 음성인식의 대상이 되며, 그 크기가 크면 인식대상이 많아지게 되므로 인식기의 성능이 낮아진다.

(4)소음환경 : 사용자가 음성인식 기술을 사용하는 환경에 따라 소음환경이 달라지게 되는데, 이러한 소음환경은 음성인식기의 성능에 큰 영향을 끼치게 된다. 조용한 환경의 밀착형 마이크는 가장 좋은 성능을 보장하지만, 일반적인 환경은 아니다. 음성인식 인터페이스를 사용하는 일반적인 환경은 (a)실내에서 고정된 마이크를 사용하는 경우와 (b)음성인식 시스템을 포함하는 모바일 단말기의 마이크를 사용하는 경우, (c)유무선 통신을 이용해서 원거리로 전송된 음성을 사용하는 경우로 나눠볼 수 있다. 위에서 언급한 세가지 경우는 모두 소음환경과 채널환경이 다르므로, 사용되는 음성인식 기술이 조금씩 다르게 된다.

2. 음성 인터페이스 관점의 음성인식 기술

일반인들이 자연스럽게 사용할 수 있는 음성인식 기술은 아니지만, 많은 응용시스템에 음성 인터페이스로서 사용되고 있다.

가장 널리 사용되고 있는 응용은 휴대전화의 음성 다이얼링이다. 음성인식 시스템을 포함하는 휴대전화기에서 전화기내에 저장된 전화번호부의 단어를 대상으로 음성인식을 수행하게 된다. 인식할 단어들 간의 특별한 문법관계는 존재하지 않으며, 음성인식 시스템의 음성모델 제작에는 사용자의 음성이 사용되지 않았으므로 화자독립이 된다[4].

텔레매틱스 단말기에서의 음성인식은 차량이라는 특수환경 때문에 고난이도의 잡음처리 기술이 요구된다. 인식대상은 단순한 명령어의 집합이며, 특별한 문법은 사용되고 있지 않지만 응용의 요구에 따라 확장될 전망이다. 휴대전화기의 음성인식 시스템과 동일하게 화자독립으로 사용되고 있다[5].

유무선 통신망 기반 음성 인터페이스로는 철도청과 KT가 개발한 철도예약시스템을 들 수 있다. 전국의 철도역명 및 명령어 등을 인식대상으로 하며, 기계주도로 대상 단어들을 인식하게 된다[6].

딕테이션 시스템은 음성인식을 통한 문서작성 시스템을 말한다. 미국에서는 Dragon Systems의 Naturally Speaking과 IBM의 Via Voice 등이 대표적인 제품이다. 국내에서는 보이스텍의 ByVoice라는 딕테이션 시스템이 출시되어 있다. 이 시스템을 사용하기 위해서는 화자 적응 과정을 거치게 되는데, 사용자는 20-30분정도 시스템에서 제시하는 문장들을 읽어서 음성 모델을 자신에게 적응시키게 된

다. 인식대상 단어는 추가가 가능하며, 분당 700타 정도의 입력이 가능하다[7].

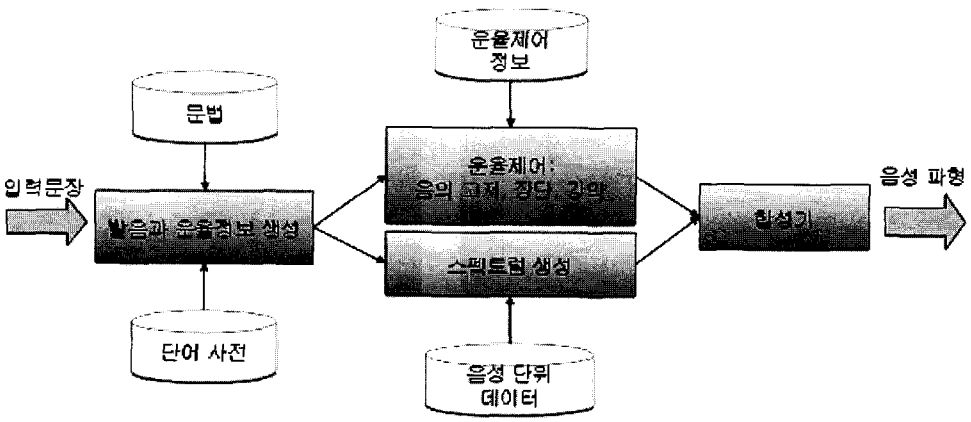
III. 음성합성 기술

음성합성이란 임의의 텍스트를 음성으로 출력하는 기술이다. 그리고, 음성합성 기술은 응용분야에 따라 다양한 방법으로 시스템을 개발할 수 있어서, 음성 인터페이스를 위한 응용 중 가장 널리 사용되고 있다.

<그림 2>는 일반적인 음성합성 시스템의 주요 구성을 나타내고 있다. 입력된 텍스트는 사전과 문법 등의 지식을 활용한 자연어처리 과정을 거치며 발음과 운율정보를 생성한다. 운율정보는 운율제어 규칙에 따라 음성의 고저, 장단, 강약 등의 운율제어 정보로 변환되고, 발음정보는 음성DB내에서 사용할 음성단위의 선택에 사용된다. 합성과정에서는 신호처리 기술을 사용하여 음높이와 음의 길이를 조절하고, 스펙트럼 보간법 등을 사용하여 음성단위사이에서 발생하는 음질저하를 방지하며 합성음을 출력한다. 최근에는 각각의 음성단위별로 좌우 음소문맥, 음높이, 음의 길이 등을 고려하여 합성할 음성단위를 선택하고 연결하여 합성음을 출력하는 코퍼스 기반 합성 방식도 많이 사용된다[8]. 이 방식은 다수의 합성단위가 미리 저장되어 있어야 하므로 많은 메모리 용량과 탐색시간이 필요하다.

음성합성에서 가장 중요한 평가요소는 합성음의 명료성과 자연성이라고 할 수 있다. 이러한 성능평가요소는 사용되는 음성합성 기술의 종류에 따라 많은 영향을 받게 된다[2].

(1) 제한된 도메인의 음성 연결방식 : 이 방법은 미리 준비된 작은 수의 음성단위를 사용하며, 주어진 도메인에서 대해 매우 높은 음질을 보장한다. 이 방법은 많은 수의 ARS시스템



〈그림 2〉 음성합성 시스템의 주요 구성[1]

에서 사용되고 있으며, 임의의 문장을 음성으로 출력하는 기능은 없다. 합성된 음성의 속도 조절이나 높낮이 조절 등도 하지 않는다.

(2) 신호처리 없는 연결방식 : 위의 방법과는 달리, 이 방법은 임의의 문장을 음성으로 출력할 수 있다. 그러나 음성단위를 잘못 연결하게 되면 음질저하가 일어나게 된다. 준비된 음성단위가 다양한 코퍼스기반 합성방식의 경우에는 음질저하의 가능성이 줄어들게 된다. CPU와 메모리 측면에서 높은 용량이 필요하며, 다양한 발성속도와 운율에 대해 음성단위가 준비되어 있으면 속도조절이나 높낮이 조절 등이 가능하다.

(3) 신호처리를 포함하는 연결방식 : 이 방법에서는 음성단위의 부드러운 연결을 위하여 음성신호가 일부 변경된다. 이 과정에서 합성된 음성의 운율이 일부분 왜곡되기도 한다. 이 방법도 신호처리 없는 연결방식과 동일하게 미리 준비된 음성단위가 있으면 합성음의 속도나 높낮이 조절이 가능하다.

(4) 규칙기반 합성방식 : 이 방법은 미리 준비된 음성단위 없이 음성을 합성하게 되므로, 전체적으로 단조로운 합성음을 생성하는 경향이 있다. 위의 연결에 비해 낮은 컴퓨팅 사양을 요구하며, 합성음의 속도조절이나 높낮이 조절이 가능하다.

음성합성 기술은 명료성이라는 성능에서 조금 부족하더라도 음성 인터페이스로 사용하는 것이 가능하다. 이에 반해, 음성인식 기술은 성능이 부족하면 오인식에 의한 사용자의 불편 때문에 음성 인터페이스로 사용하는 것이 불가능하다. 이러한 이유로 음성인식 기술보다는 음성합성 기술이 오래전부터 음성 인터페이스로 많이 사용되어 왔다.

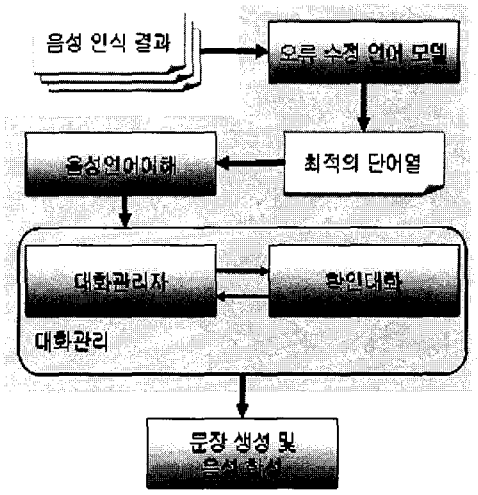
현재 음성 인터페이스로 널리 사용되고 있는 음성합성 기술은 제한된 도메인의 음성연결방식이며, 응용 시스템에 따라서 합성단위를 새로이 구축하고 있다. 철도예약시스템 등과 같이 미리 정해진 음성대화가 필요한 경우에 유용하다.

무제한 도메인에 대한 음성합성이 필요한 응용시스템의 경우에는 신호처리가 없는 연결방식을 사용하는 기술이 많이 사용되고 있다. 정부기관의 홈페이지에서 장애우를 위해 TTS(Text-to-Speech) 서비스를 제공하는 시스템에서 사용하고 있는 기술이다.

Ⅳ. 음성 인터페이스 기술의 미래

아직도 음성인식 기술과 음성합성 기술은 독립적인 인터페이스 기술로 사용하기에는 부족한 점이 있다. 하지만, 기술의 발달과 컴퓨팅 성능의 개선으로 사람과 컴퓨터가 음성 인터페이스를 사용하여 자연스럽게 대화하는 기술이 필수적인 지능형 인터페이스로 연구되고 있다. 이런 환경에서 음성대화관리기법을 사용하면 사용자는 보다 완벽하고 편리한 음성 인터페이스를 제공받을 수 있다. 각각의 응용시스템을 위한 음성대화관리기법은 <그림 3>의 프레임워크에 따라 구현된다.

[9]에서의 음성대화관리시스템은 일반적으로 음성인식(Speech Recognition) 및 언어모델적응(Language Model Adaptation), 음성언어이해(Spoken Language Understanding), 대화관리



〈그림 3〉 대화형 음성 인터페이스의 구조

(Dialogue Management) 및 확인대화(Clarification and Verification), 그리고 음성합성(Text-To-Speech Synthesis) 으로 이루어져 있다.

1. 언어모델적응

음성인식 기술의 낮은 인식률은 인터페이스로서의 효용성을 떨어뜨리게 된다. 이러한 낮은 인식률은 학습 데이터와 인식 데이터가 서로 다를 때 특히 두드러지며, 이를 해결하기 위해서는 음성인식기를 인식할 대상 도메인에 적응시키는 과정이 필요하다.

오류수정 언어모델(Error Corrective Language Model) 적응 기법에서는 적응 데이터를 채널 적응을 위한 음성인식 결과와 언어자질을 포함한 텍스트로 나눈다. 여기서 모델 적응 문제는 입력 단어열에서 확률 $Pr(w'|w)$을 최대화시키는 최적의 수정된 단어열 w'를 찾는 과정으로 볼 수 있다.

이 모델은 채널모델과 언어모델로 구성되며, 잡음채널 모델을 따른다. 이는 최대 후험확률 (Maximum A Posteriori) 적응 혹은 오류 최소화를 위한 최적화 모델로 해석된다. 채널모델은 인식 환경의 채널 특성을 반영하며 이를 통해 언어 모델이 도메인, 잡음환경, 또는 화자에 적응이 된다.

2. 음성언어이해

음성언어이해는 음성인식의 결과를 분석하여 대화관리모듈이 처리할 수 있는 형태의 의미구조로 변환하는 역할을 수행한다. 음성인식의 한계 때문에 나타나는 여러 가지 제약 조건을 극복하기 위해서는, 미리 정의된 의미구조의 필수적인 요소만 추출하는 것을 목적으로 하는 개념 집어내기(Concept Spotting) 방식 혹은 시퀀스 레이블링(Sequential Labeling ; CRF) 방식의 언어 이해가 가장 적합하다. 이러한 방식은 미리 정의된 의미구조인 슬롯(slot)만을 처리하기 때문에, 언어를 부분적으로 이해하는 방식이다. 하지만, 특정 영역의 언어이해를 위해 슬롯을 적절하게 설계한다면, 각 슬롯의 값으로부터 언어이해에 필수적인 정보를 얻을 수 있다는 장점이 있다.

3. 대화관리

대화관리는 음성 인터페이스를 위한 대화시스템에서 중추적인 역할을 하는 부분으로, 사용자의 음성을 받아 의미를 추출하고 사용자에게 필요한 정보를 제공하기 위해 외부 지식자원들을 연결하여 시스템 발화를 생성하는

전반적인 대화 흐름을 제어하는 부분이다.

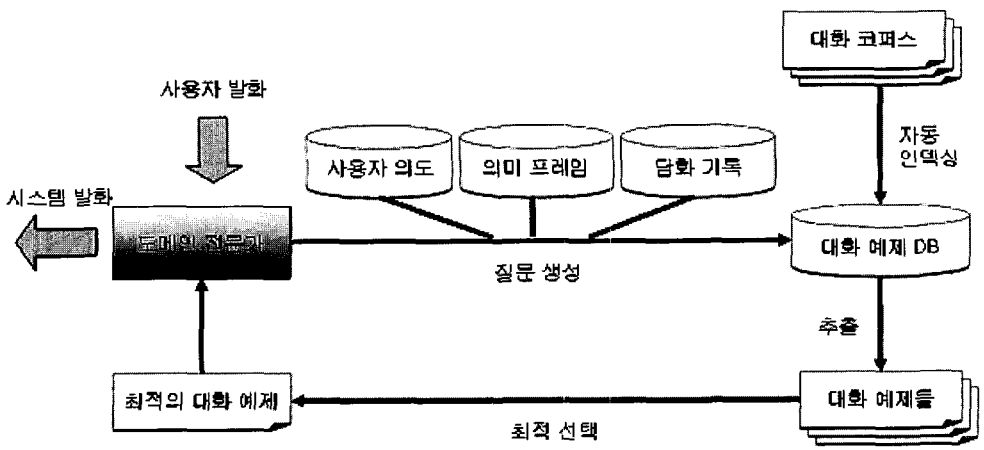
대화 예제를 이용한 상황 기반 대화관리 시스템의 기본적인 아이디어는 프레임 기반의 대화 모델을 취한다는 점이다. 그리고, 대화 예제를 통하여 규칙을 자동으로 학습하여 도메인 확장성을 증대시킨다.

상황기반 대화관리라는 것은 특정한 상태 전이 규칙에 얽매이지 않고 현재의 대화 상황에 따라서 어느 상태로든지 전이될 수 있다는 것을 의미한다. 여기서 대화 상황이라는 것은 사용자의 의도, 현재 프레임의 내용, 그리고 슬롯이 채워진 정도에 대한 담화 기록 등을 이용하여 정의한다. 이것은 특정한 문법이나 상태 전이 없이 좀 더 자유로운 사용자의 발화를 관리할 수 있으며 도메인에 대한 특성화된 방법이 없어서 새로운 도메인에 적용하기가 쉬운 장점이 있다.

대화 예제를 이용한 대화모델링은 대화 말뭉치로부터 현재 대화 상황에 맞는 대화 예제를 찾아서 적절한 시스템 발화를 하기 때문에, 규칙기반 모델에서 필요한 규칙 코딩 작업을 최소화할 수 있다. 준비된 대화 말뭉치로부터 자동적으로 규칙을 학습하여 대화 예제 데이터베이스를 만들고 그것을 검색하여 시스템 응답을 찾는 방식이다. 따라서, 대화 예제 기반 모델링은 규칙 기반 대화 모델링의 단점인 규칙 학습의 인력 비용을 절감할 수 있다. 이로 인해 대화관리 시스템의 도메인 확장이 효율적이고 손쉽게 되었다. <그림 4>는 대화 예제 기반의 대화모델링의 전반적인 전략을 도식적으로 보여준다.

4. 확인대화

확인대화 (Clarification Dialogue)란 사람과 사람 혹은 사람과 컴퓨터 사이의 대화에서 적절치 못한 정보가 전달된 경우, 즉 불분명 혹은 불충분하거나 혼동의 여지가 많은 정보를



〈그림 4〉 대화 예제 기반 대화모델링

대화를 통해서 재확인해나가는 대화 유형을 말한다.

이러한 확인대화를 위하여 3단계 정보 확인 방법에서는 음성인식단위 확인, 정보보존율에 의한 문장 확인, 대화정보 확인을 하게 된다. 우선 음성인식단위 확인에서의 오류검출이 이루어지고 이러한 정보가 다음 단계인 문장 확인으로 전달된다. 문장 확인에서는 정보보존율에 의거하여 음성인식과 음성언어이해 전체에 걸쳐 그 문장이 어느 정도 믿을 만한지를 판단한다. 이러한 정보 보존율이 다음 단계인 대화정보 확인으로 건네지게 되어 대화 진행에 필요한 대화정보단위에서 그 정보가 음성인식 측면에서나 이해 측면, 그리고 관계 측면에서 얼마나 잘 전달되었는지를 검증하게 된다.

V. 결론

본 고에서는 음성 인터페이스를 위한 기본적인 음성처리 기술을 설명하고, 그 응용에 필요한 고려사항들을 살펴보았다. 또한, 미래의 음성 인터페이스로 활용될 대화 관리시스템의 개념에 대해서도 알아보았다.

음성인식 기술과 음성합성 기술은 많은 연구를 거쳐 다양한 응용 시스템에 음성 인터페이스로 사용되고 있으며, 그 활용도는 점점 넓어져 갈 것이다. 하지만, 기존의 응용 시스템에 현재의 음성처리 기술을 인터페이스로 사용하려는 노력과 더불어, 현재의 음성처리 기술의 한계를 알고, 그것에 최적화된 응용시스템을 개발하는 것이 음성 인터페이스의 사용을 넓히는 길이 될 것이다. 물론 음성처리 기반기술의 성능을 높이는 연구도 지속적으로 수행되어야 할 부분이다.

공상 과학 영화에 나오듯이 인간처럼 대화할 수 있는 음성대화 인터페이스가 짧은 시간 안에 출현하지는 않겠지만, 사람과 기계의 궁극적인 인터페이스 수단으로 지속적인 연구 대상이 되어야 할 것이다. 음성대화 인터페이스를 위한 연구로서 음성인식 및 언어모델적응, 음성언어이해, 대화관리 및 확인대화, 그리고 음성합성으로 이루어진 대화관리 시스템을 제시하였으며, 이러한 연구로 발전된 음성 인터페이스 기술이 각종 상품에 적용될 수 있기를 기대해 본다.

참고문헌

[1] R. Cole et al., Ed., Survey of the State of the Art in Human Language Technology, Cambridge University Press, 1997.

[2] M. Jiang, T. Schultz, Paper #3 Review of SpeechTechnologies, http://www.ells.edu.cn/actives/zjzhy/blw/3.Speech.tech-jiang.doc

[3] 김형순, 음성언어정보처리기술의 현황과 전망, 대한전자공학회지, 대한전자공학회지, 2003, 30(7), pp.700-708

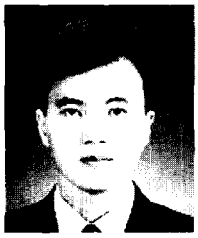
[4] 구명완, 김재인, 음성정보처리기술 응용서비스, 한국정보처리학회지, 2004, 11(2), pp.17-24

[5] 정민화, 자동차용 음성 HMI 시스템 기술개발, 한국정보처리학회지, 2004, 11(2), pp.42-47

[6] 심유진, 김재인, 구명완, 음성인식을 이용한 자동 호 분류 철도 예약 시스템, 대한음성학회지:말소리, 2004, (52), pp.161-169

[7] 김정인, 제품 소개2 : 한글 받아쓰기 S/W 바이보이스(ByVoice) 소개, 한국멀티미디어학회지, 7권, 2호, pp.93-95
[8] 김회린, 음성정보처리 기술 개발 현황 및 전망, 한국정보처리학회지, 2004, 11(2), pp.25-32
[9] 정민우, 은지현, 이청재, 정상근, 이근배, 음성 자연어 처리를 위한 대화 관리 시스템, 한국정보과학회지, 2006, 24(1,200), pp.19-26

저자소개



김 병 창

1995년 2월 경북대학교 컴퓨터공학과(학사)
1997년 2월 포항공과대학교 컴퓨터공학과(석사)
2002년 2월 포항공과대학교 컴퓨터공학과(박사)
2002년 3월-2004년 8월 위덕대학교 전임강사
2004년 9월-현재 대구가톨릭대학교 조교수

주관심 분야 : 음성언어정보처리, 임베디드시스템

저자소개

정 민 우

2003년 2월 전북대학교 컴퓨터공학과(학사)
2005년 2월 포항공과대학교 컴퓨터공학과(석사)
현재 포항공과대학교 컴퓨터공학과 박사과정

주관심 분야 : 언어 모델, 언어 이해

이 근 배

1984년 2월 서울대학교 컴퓨터공학과 학사
1986년 2월 서울대학교 컴퓨터공학과 학사
1991년 UCLA 컴퓨터학과 박사
1991년 3월-1991년 9월 UCLA 연구원
1991년 9월-1997년 2월 포항공과대학교 조교수
1997년 3월-2004년 2월 포항공과대학교 부교수
2000년 9월-2001년 8월 미국 Stanford CSLI 연구원
2004년 3월-현재 포항공과대학교 정교수

주관심 분야 : 자연언어 처리, 음성인식, 정보검색, 바이오 인포메틱스