# 자연어 음성인식 기술을 이용한 음성 대화 서비스 개발동향

## Spoken Dialogue Service Trends Using Natural Speech Recognition Technology

소프트웨어 기술의 미래전망 특집

정호영 (H.Y. Jung) 음성처리연구팀 책임연구원
송화전 (H.J. Song) 음성처리연구팀 선임연구원
강병옥 (B.O. Kang) 음성처리연구팀 선임연구원
정의석 (E.S. Chung) 음성처리연구팀 선임연구원
정 훈 (H. Chung) 음성처리연구팀 선임연구원
오유리 (Y.R. Oh) 음성처리연구팀 연구원
권오욱 (O.W. Kwon) 언어처리연구팀 책임연구원
이기영 (K.Y. Lee) 언어처리연구팀 선임연구원
이윤근 (Y.K. Lee) 음성처리연구팀 팀장

목 차



모바일 혁명과 빅데이터(big data) 시대에 접어들면서 사용자 중심의 자연스러운 인터페이스와 정보 검색에 대한 요구가 늘어가고 있다. 모바일 환경에서의 쉽고 자연스러운 검색을 위해 음성인식 기술을 이용한 음성검색 서비스가 대세를 이루고 있으며 대화형 검색 서비스로 발전하게 될 것이다. 음성 대화 서비스의 주요 응용 분야인 음성검색 및 외국어 교육 서비스에서의 자연어 음성인식 기술 역할 및 사용자 경험을 바탕으로 하는 선순환 구조의 인식 성능 개선에 대해 소개한다. 또한 두 응용분야에서의 국내외 개발동향을 소개하고 실제 개발 사례를 통해 무제한급 자연어 음성인식 기술에 기반한 음성 대화 서비스의 가능성을 살펴본다.

## Ⅰ. 서론

모바일 기술 및 클라우드 시스템의 성장으로 대부분의 IT 연구기관 및 업체들이 편리한 인터페이스 및 효과적인 정보검색에 관심을 가지고 있다. 이에 음성대화 시스템을 이용한 자연스러운 인터페이스 및 대화형 정보검색 서비스의 요구가 증가하고 있는 상황이다. 대표적 검색업체인 구글도 모바일 검색 서비스를 제공하면서 음성검색 기능을 필수적으로 제공하고 있으며 사용자 데이터를 이용하여 사람이 수행하는 수준으로 발전시키려 하고 있다. 이것은 무제한급 자연어 음성인식 기술을 기반으로 하여 음성 대화 서비스를 제공하는 것을 의미한다.

국내에서도 다음을 시작으로 네이버가 음성검색 서비스를 제공하고 있으며, 구글의 발표에 의하면 모바일 검색 가운데 음성검색의 비율이 20%에 달하고 있으며 이 비율은 더 증가할 것으로 예상되고 있다. 음성검색 기술은 무제한급 자연어 음성인식 기술을 필요로 하며, 음성 대화 서비스의 출발선으로 볼 수 있다. 유선 인터넷 검색에 비해 모바일 검색의 차이점은 정확한 검색이 되어야 하고 개인 최적화 검색이 요구된다는 점이다. 이것을 만족시키기 위해서는 단순 음성검색만을 제공하는 것이 아니라 대화를 통해 사용자 최적화된 정확한 검색을 제공하는 음성 대화 서비스가 요구된다. 따라서 모바일 환경에서의 음성대화 서비스는 주목할 가치가 있으며, 이의 핵심인 자연어 음성인식 기술의 성능 개선이 요구되고 있다.

모바일 환경에서의 편리한 정보 검색뿐만 아니라 음성 대화 서비스는 외국어 학습을 위해 큰 관심을 받고 있다. 시간 및 공간의 제약에 비교적 자유로우며 학습자가 실수에 대한 부담을 적게 가져 외국어 학습에 음성 대화 서비스를 응용하려는 많은 시도가 있어 왔다. 특히 외국어 말하기 능력 향상을 위해 대화형 학습에 대한 요구가 크게 증가하고 있는 상황이다. 정보검색과 달리 외국어 학습을 위한 대화관리를 하는 점에서 차이가 있지만, 결국 자연어 음성인식 기술이 대화형 학습의 핵심으로 볼 수 있다. 단순 음성검색에서 대화형 검색 서비스로 발전하는 것처럼 외국어 발음 평가에서 대화형 말하기 교육으로 발전하게 될 것이며, 이것은 자연어 음성인식 기술에 바탕을 둔 음성대화 서비스를 통해 이루어질 수 있을 것이다.

본 고에서는 음성 대화 서비스에서의 자연어 음성인식 기술의 역할을 소개하고 사용자 경험 및 로그데이터를 바탕으로 인식 성능을 지속적으로 개선하는 선순환 구조를 소개하며, 음성검색 서비스 및 대화형 영어교육 서비스를 통해 음성 대화 서비스의 필요성을 살펴보고자 한다.

## Ⅱ. 음성 대화 서비스에서의 자연어 음성인식 기술

음성인식 기술은 간단한 명령어 인식에서 시작하여 낭독체 연속어 인식, 대규모 연속어 인식을 거쳐 무제한급의 자연어 인식의 단계로 발전하고 있다. 2007년 이후 다양한 정보에 편리하게 접근하기 위한 대화 시스템의 필요성과 함께 대화 시스템의 출발선이라 할 수 있는 모바일 환경에서의 빠른 정보검색을 위한 음성검색 시스템이 개발되어 왔다. 이와 같은 서비스를 위해서는 자연어 음성인식 기술이 요구되며 시대적 요구에 맞게 음성인식 기술도 발전하고 있다. 특히 2007년 이전이 음성인식의 본질적인 어려움을 해결하기 위한 방법들이 개발된 시기라면, 2007년 이후는 음성인식에 있어 해결하지 못한 문제를 다양한 지식을 활용하여 개선하려는 시기하고 볼 수 있다. 음성인식 기술의 보편적 활용에 있어 가장 큰 문제점은 사용자에 따른 인식률의 차이, 주변 잡음에 따른

인식률 저하, 인식대상 어휘의 제한으로 인한 인식오류 발생으로 볼 수 있다. 이 문제들을 해결하기 위해 다양한 지식을 활용하는 음성인식 프레임의 필요성이 증가하고 있다.

음성인식을 위한 다양한 지식은 acoustic channel 및 linguistic channel의 두 관점에서 볼 수 있다. acoustic channel 에서는 화자, 반향상태, 배경잡음, 마이크로폰 등의 지식을 활용할 수 있고 linguistic channel 에서는 언어, 어휘, 문법, 문맥 등의 지식을 활용할 수 있다. 음성인식 상황에서 두 관점의 지식을 메타데이터로 표현할 수 있고 지식을 통계적 음성인식 프레임에 결합하는 방법론이 있다면 음성인식의 성능은 크게 개선될 수 있으며, 자연어 음성인식의 최종 목표인 누구나(anybody), 어디서든(anywhere), 어떤 말이라도(anything) 인식할 수 있는 시스템을 얻을 수 있을 것이다.

최근 들어 스마트폰을 중심으로 한 모바일 인터넷 환경과 클라우드 서비스의 확대로 향후 웹 서비스에 큰 변화를 가져올 5대 요소로 mobile, video, sensors, big data, natural interface가 선정되었다[1]. 즉, 모바일 환경으로 인해 자연스러운 인터페이스의 수요가 확대되고 이로 인해 많은 사람들이 음성인식을 사용함으로써 엄청난 규모의 사용자 로그 데이터를 확보할 수 있다는 것이다. 이것은 다양한 배경환경에서의 다양한 화자가 말한 다양한 어휘를 확보할 수 있는 것을 의미하며, anybody, anywhere, anything 음성인식을 위한 발판을 마련할 수 있는 계기가 되고 있다. 실제 구글은 음성검색 서비스를 통해 하루 동안 한 사람이 2년 동안 쉬지 않고 얘기하는 양의 음성데이터를 수집하고 있으며, 음성검색 시스템의 개발 책임자인 마이크 코언은 스마트폰과 PC 등 기기의 장벽을 넘어선 음성인식이 가능해지고 있는데 이것은 사람들이 돌아다니는 곳 어디서든 기계가 사람의 음성에 반응해 움직이는 시대의 시작을 의미한다고 주장하였다. 다양한 환경에서 다양한 화자가 발성한 다양한 어휘, 문법 등을 분석함으로써 무제한급 자연어 음성인식을 위한 acoustic 및 linguistic 지식을 체계화 할 수 있다. 따라서 모바일 환경 음성검색 서비스는 자연어 음성인식 기술을 필요로 할 뿐 아니라 자연어 음성인식 기술을 개선할 수 있는 선순환 구조에 있다고 볼 수 있다.

편리한 정보 접근 외에 음성 대화 시스템을 활용할 수 있는 응용 분야는 외국어 학습에 관한 것이다. 외국어 학습의 경우 대화를 통해 자주 접하는 것이 큰 도움이 되는 것은 의심할 여지가 없다. 그러나 전문 외국어 강사 또는 원어민 강사의 부족과 학습자의 흥미를 유발한 학습 방법의 부족이 문제가 되고 있다. 이런 문제를 해결하기 위해 음성 대화 시스템을 이용해 외국어 교육을 시도하려는 움직임이 계속되고 있다. 비대화형 외국어 교육에 비해 대화형 외국어 교육은 시나리오 및 역할 부여를 통해 학습자의 흥미를 유발할 수 있고 시간 및 공간적 제약에서 비교적 자유로우며 강사와의 교육에 비해 실수에 대한 부담이 적은 장점을 가지게 된다. 최근 국내에서도 국가영어능력평가시험인 NEAT를 도입하려고 하고 있으며, 말하기 능력 평가에 대비한 교육방법이 절실히 요구되고 있다.

대화형 외국어 교육을 위해서는 자연어 음성인식 기술, 언어 이해 기술 및 대화 생성 기술이 요구된다. 다른 기술도 대화형 학습을 위해 중요하지만, 학습 효과를 최대로 하기 위해서는 자연어 음성인식 기술 및 오류 검출 기술이 가장 기본이 됨은 틀림없다. 음성검색 서비스와 마찬가지로 대화형 외국어 교육 서비스를 통해 학습 단계별 사용자의 발성 데이터를 수집할 수 있으며 이는 자연어 음성인식의 성능을 개선하는 선순환 구조를 이룰 수 있다. 또한 학습자 수준별 다

양한 데이터를 수집함으로써 오류를 범하기 쉬운 발음 및 구문 등에 대한 지식을 체계화할 수 있으며, 이로 인해 자연어 음성인식을 위해 지식을 결합하는 플랫폼에 활용할 수 있어 인식 오류 검출, 언어 이해 등의 성능을 개선하여 대화 시스템 전체의 성능을 개선하는 효과를 가져올 수 있다.

## Ⅲ. 음성검색 서비스

### 1. 배경 및 음성검색 서비스 동향

#### 가. 배경

음성검색 서비스는 키보드를 통한 문자입력이 어려운 상황에서 음성인식 기술을 이용하여 핵심어를 입력하여 이와 관련된 다양한 정보를 제공한다. 이러한 음성검색 서비스는 일반적으로 전화망에서의 자동 응답 서비스의 입력 수단으로 원하는 서비스와의 연결 또한 주가 조회 등에 음성인식 기술을 사용한 것이 시초라고 할 수 있으며, 현재는 모바일 환경의 스마트폰을 통해 이전에는 상상할 수 없었던 다양한 서비스를 선보이고 있다.

#### 나. 음성검색 서비스 동향

현재 음성검색 서비스를 주도하는 업체는 구글이며, MS(Microsoft)의 경우에도 구글과 거의 비슷한 개발 역사를 가지고 있다. <표 1>에 구글과 MS에서 지금까지 개발한 음성검색을 이용한 서비스를 정리하였다. 특히 구글의 경우 스마트폰에서 활발하게 사용되는 음성검색 앱을 개발한 2008년 이후 현재까지 총 13개 국어에 대한 음성검색 서비스를 제공하고 있다.

MS의 경우도 윈도폰 7(Windows Phone 7)과 X박스 등에 자체 검색엔진인 Bing과 함께 음성검색 기능을 추가하였으며, 또한 아이패드용으로도 앱을 개

〈표 1〉 구글과 MS의 음성 검색 서비스 개발 내역

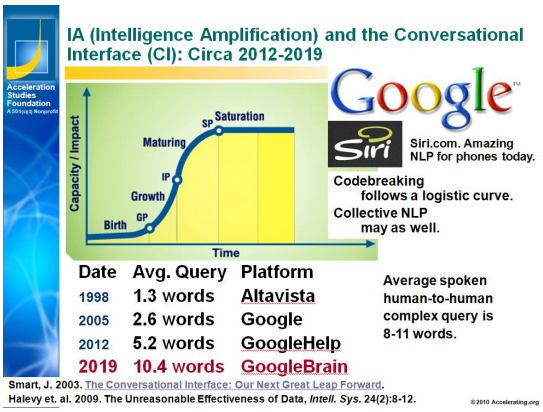
| | | |
|---|---|---|
| 구글 | 2007. 4. | GOOG-411 |
| | 2008. 7. | Google Maps with voice search |
| | 2008. 11. | Google mobile App with voice search |
| | 2010. 3. | Google voice search in Youtube |
| | 2011. 6. | Google voice search on Google.com |
| MS | 2007. 10. | CALL-411 - MS voice search |
| | 2009. 4. | Tellme service in MS Window Mobile 6.5 |
| | 2010. 10. | Bing voice search - Windows Phone 7 |
| | 2011. 4. | Bing app with voice search for iPad |
| | 2011. 6. | Bing voice search on Xbox |

발하여 iOS에서의 음성검색엔진을 제공하여 모바일 시장에서의 주도권을 잡기 위한 행보를 진행하고 있다.

지금까지 음성 기술분야의 선두는 뉘앙스였다. 그러나, 클라우드 컴퓨팅 기반으로 IT 생태계가 급변하는 동안 이에 대한 대응이 늦어짐으로써 모바일 시장에서는 구글에 뒤쳐지게 되었다. 또한 애플의 경우도 iOS가 안드로이드에 비해 음성검색과 음성 입력 기능이 없는 관계로 이 부분에 대해서는 안드로이드 진영에 밀리는 형상이나 최근 뉘앙스에서 Dragon Search 및 Diction 앱을 iOS 용으로 개발함으로써 어느 정도는 안드로이드의 음성검색 서비스에 대항하게 되었다.

구글, MS 및 뉘앙스 외에 기존의 음성기술 강자였던 IBM과 AT&T는 자체 음성검색 솔루션(예: Speak-4it-AT&T) 개발뿐만 아니라 구글이나 MS와는 다르게 현재 IT 시장에서 진입 장벽이 높은 음성 기술을 SaaS(Speech As A Speech) 형식의 서비스를 통해 지금까지 개발한 최첨단 음성인식엔진 및 플랫폼을 음성 기술을 통해 부가가치를 창출하고자 하는 기업들에게 제공하여 새로운 생태계 창출을 도모하고 있다.

또한 국내의 경우에는 주요 포털 사이트인 다음과




(그림 1) 구글 음성기술 향후 개발 계획

네이버에서 음성검색 기술을 모바일 서비스에 제공하고 있다. 다음의 경우는 ETRI와 기술 제휴를 통해 2010년 6월에 국내 최초로 음성검색 서비스를 상용화하였고, 현재 이를 기반으로 지도검색, 음악검색, 쇼핑 등의 영역으로 음성검색 기술을 확대 적용하고 있다. 또한 다음에 이어서 구글에서 한국어 음성검색 서비스를 개설하였고, 2010년 12월에 네이버도 본격적으로 음성검색 서비스를 제공하였다.

이러한 음성검색 서비스의 향후 발전 방향은 (그림 1)의 구글의 향후 개발 계획에 잘 나타나 있다. 사람의 경우 질문을 위해 평균적으로 8~11 단어로 구성된 문장을 사용한다. 현재 구글의 경우는 5 단어 정도의 문장을 인식하는 수준이다. 이를 확장하여 2019년경에는 거의 사람이 수행하는 정도의 수준(Google-Brain)으로 발전시키고자 한다. 이는 모바일 환경에서 엄청난 수의 안드로이드폰 사용자들을 통해 전세계 각지에서 감당하기 힘든 정도의 엄청난 양의 다양한 언어의 음성 데이터를 모을 수 있으며, 이를 통해 수집된 음성을 가공하여 성능 개선에 활용하는 선순환 구조를 형성하였기 때문에 가능할 수 있다. 이러한 빅데이터를 이용한 서비스 품질의 개선은 현재 모든 IT 서비스 영역에서 발생하고 있는 새로운 생태계이다. 또한 이름에서 알 수 있듯이 음성인식의 수준을 넘어서 음성 이해의 수준으로 진화해 사용자의 말을 이해하고 그 의도를 판단하여 그에 해당하는 서비스까지도 제공해 줄 수 있을 것이다.

## 2. 음성검색 서비스를 위한 요소 기술

본 절에서는 무제한급 자연어 음성인식을 위한 요소기술을 음성검색 서비스와 연관시켜 소개한다. 음성검색 서비스를 통해 음성 빅데이터를 얻을 수 있고 이를 음성인식 성능 개선에 활용하는 선순환 구조를 이룰 수 있게 되는데, 이런 빅데이터를 이용해 음향 및 언어의 통계적 모델을 향상시키는 방법을 소개한다. 또한 무제한급 어휘의 음성인식을 실시간 속도로 처리하기 위해 많은 수의 프로세서를 활용하는 병렬처리 음성인식에 대해 기술한다.

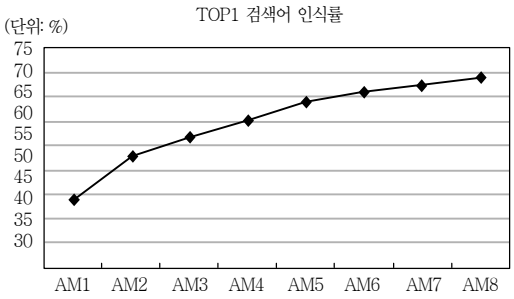
### 가. 로그 데이터기반 음향모델링

음향모델링 기술은 음성인식을 위해 불특정 다수 화자의 다양한 발음특성을 모델링하는 것을 목적으로, 대용량의 음성데이터로부터 통계적 방식으로 모델 파라미터 형태로 표현되는 참조패턴을 생성하는 기술이다. 음성인식 시스템이 좋은 성능을 갖기 위해서는 다양한 환경, 화자, 어휘로부터 얻어진 대용량 훈련데이터로 훈련된 음향모델이 필요하다. 특히 훈련용 음성데이터가 분포하는 환경이 실제 음성인식 시스템이 사용되는 환경에 가까울수록 훈련환경과 실제 사용환경의 불일치를 줄임으로써 높은 수준의 음성인식 성능을 보장할 수 있다.

여기서 실제 사용환경은 음성인식 시스템이 사용되는 채널환경, 사용자 주변환경, 사용자 발화패턴 등을 포함하는 acoustic channel 지식의 개념으로, 음성인식 서비스를 이용하는 사용자로부터 얻어진 음성로그는 실제 사용환경을 가장 잘 반영하는 훈련데이터가 될 수 있다. ETRI에서 개발한 음성검색 시스

템에서도 다양한 환경, 화자, 음성분포를 갖는 대용량 음성데이터를 기반으로 모바일 음성검색 서비스를 통해 수집된 실제 사용자 음성로그를 단계적으로 더하여 음성검색 서비스의 성능을 점진적으로 향상시킬 수 있었다.

음성검색 시스템의 개발 초기에 있어서는 음향모델을 훈련하기 위해 일반적인 영역에서의 다양한 음성데이터를 활용하게 된다. 예를 들어 음소의 균형을 맞춘 고립단어 데이터, 다양한 화자 및 검색어의 특성을 반영하기 위한 저빈도 음소 데이터, 성인 및 아동 발성 데이터 및 다양한 단말 특성이 반영된 음성데이터를 사용할 수 있다. 이런 데이터는 주로 다양한 음소적 특성, 화자 특성 등을 반영하기 위한 기본 데이터로서 다양한 잡음이 반영되는 이동환경에서 주로 사용하는 모바일 음성검색 서비스에서는 성능저하를 피할 수 없다. 이를 해결하기 위한 데이터를 음성검색 서비스를 통한 음성 로그로부터 얻을 수 있다. 실제 사용자의 음성 로그 데이터를 대상으로 순차적으로 전사작업을 수행하는 한편, 일부는 미전사 음성 로그 데이터와 인식결과로부터 발화검증을 통해 자동으로 훈련데이터를 추출하여 음향모델에 반영할 수 있다. 음성 빅데이터를 대상으로 음향모델을 위해서는 훈련 가능한 데이터를 선별하는 기능, 미전사 데이터를 훈련에 적용하는 기술 등이 요구되며, 일부 그룹에서는 unsupervised training이라는 기술을 시도하고 있다[2]. ETRI에서 개발한 음성검색 서비스를 대상으로 로그 데이터의 활용 측면에서의 음성인식 성능 향상 추이는 아래 (그림 2)과 같다.




(그림 2) 음성 로그 활용에 따른 성능추이

### 나. 대규모 코퍼스기반 언어모델링

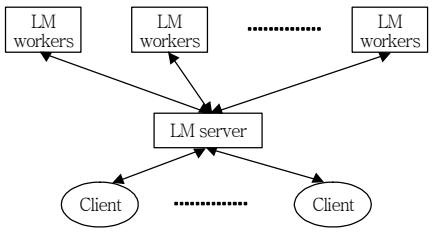
다음, 네이버, 구글 등과 같이 현재 상용화된 음성검색 시스템의 경우, 기초적인 음성 받아쓰기 기능을 제공한다. 해당 기능을 제공하기 위해서는 대규모 코퍼스기반 언어모델링 기술이 필수적이다. 웹 검색 도메인의 경우, 대상 코퍼스와 구성 어휘의 수는 기하급수적으로 증가하며, 특정 도메인으로 대상 영역을 한정할 수 없는 특징을 갖는다. 이는 언어모델의 대용량화와 지속적 확장 기능을 요구한다. 이를 해결할 수 있는 언어모델링 기술의 최근 동향은 대용량 분산 언어모델링 기술 방식이 있다.

대용량 분산 언어모델링 기술의 특징으로는 n-gram 차수의 무제한과 어휘 수의 무제한이 있다. 이는 대용량 학습 데이터를 전제로 하고 있으며, 언어모델링의 기존 smoothing 기술 의존성을 벗어날 수 있는 접근 방법이다. <표 2>는 [3]에서 보고된 자료로, target 도메인의 경우, 200k의 어휘 수와 2G 분량의 언어모델 크기를 갖는 반면, 이를 웹 영역으로 확장했을 경우, 언어모델 크기는 1.8T에 다다른다. 이는 단일 서버로 처리할 수 있는 한계를 넘어서는 분량이다. [4]에서는 분산 언어모델링 기술의 기본 구조를 소개하고 있다.

(그림 3)은 [4]의 분산 언어모델링 아키텍처를 기술한다. 여기서, LM worker는 분산되어 있는 코퍼

〈표 2〉 언어모델 크기

| | Target | Web |
|---|---|---|
| 어휘 크기 | 200k | 16M |
| n-gram 개수 | 257M | 300G |
| 언어모델 크기 | 2G | 1.8T |




(그림 3) 분산 언어모델링 구조

스 단위로 볼 수 있고, Suffix Array 구조로 구성되어 있어, 클라이언트의 n-gram 빈도수 요구 요청에 대응할 수 있다. 클라이언트와 LM server는 TCP/IP 소켓으로 구현되어 있고, server와 worker는 MPI (Message Passing Interface)로 구성되어 있다. 해당 모델은 음성인식의 N-best 리스코어링에 적용되어 5.3%의 WER 성능 개선을 보였다.

구글의 경우 자체적인 클러스터링 기술을 이용하여, 무제한의 어휘 수와 n-gram 개수를 기반으로 하여 언어모델을 제공하고 있다. 해당 클러스터링 기술은 MapReduce라는 프로그래밍 모델과 해당 라이브러리로 대용량 데이터의 처리를 목적으로 개발되었다[5]. 이는 병렬/분산 처리 프로그래밍 경험이 없는 개발자도 분산 리소스를 활용할 수 있게 한다. 구글의 MapReduce 라이브러리는 비공개되어 있으나, 오픈소스 영역의 Hadoop를 이용하여 해당 기능을 활용할 수 있다. [3]에서는 자동 기계번역 영역에서 Stupid Backoff라는 Smoothing 모델을 선보이고, MapReduce 방식에 기반한 분산 언어 모델 기술을 소개하였다. 여기서 Stupid Backoff의 경우 코퍼스에 발생하지 않은 n-gram에 대하여 smoothing 시 경험적 상수 값을 할당하는 방식을 취하며, 대용량 코퍼스 실험에서 Kneser-Ney 성능과 차이가 없음을 보였다. 실험은 2조 개의 토큰 분량의 학습 데이터와 3,000억 개 분량의 5-gram 언어모델 구축을 기반으로 진행되었다. 결론적으로, 대규모 언어모델링을 위해서는 무제한의 언어 데이터 수집 기술에 기반하여 분산 리소스를 활용한 언어모델링 기술 확보가 필수적이라고 할 수 있다.

### 다. 병렬처리 디코딩 기술

연속 분포 은닉 마코프 모델 기반의 음성인식 시스템에서 디코딩 과정이란 비터비 디코딩 알고리즘으로 정의되며 따라서 비터비 디코딩 연산 복잡도를 줄이기 위한 고속화 알고리즘은 확률모델의 계산을 효율적으로 수행하거나 계산할 확률모델의 개수를 제한하는 방식으로 구분된다. 그러나, 이런 고속화 알고리즘은 성능과 속도 간의 trade-off 관계가 있어 성능 저하를 최소로 하면서 속도 개선을 하기 위해서는 정교한 튜닝 작업을 필요로 한다[6].

최근 들어, 고성능 CPU 사용이 보편화 되면서 성능 저하 없이 고속 디코딩을 구현하기 위한 방법으로 병렬처리 기술을 사용하는 방식들이 제안되고 있다. 대표적인 방법들로 SIMD(Single Instruction Multiple Data) 명령어를 사용하는 방식, CPU 내에 포함된 여러 개의 연산 코어를 사용하는 멀티 코어(Multi-core) 방식 혹은 GPU(Graphic Processing Unit) 내에 포함된 수십 혹은 수백 개의 연산 코어를 사용하는 매니 코어(Many-core) 방식이 있다[7]. SIMD는 단일 명령어로 다중 데이터를 동시에 처리하는 병렬처리 방식으로 SSE(Streaming SIMD Extension) 명령어의 경우에는 128비트 레지스터를 사용해 4개의 단일 정밀도 부동 소수점 데이터를 동시에 처리하는 것이 가능하다. 멀티 코어 방식은 OpenMP(Open Multi-Processing) 라이브러리를 사용해 직렬 연산 작업을 멀티 코어로 분배함으로써 수행 시간을 단축시킨다. 매니 코어 방식에서는 GPU 벤더가 제공하는 SDK나 크로노스 그룹에서 제안한 OpenCL 라이브러리를 사용해 GPU 내의 다중 코어에 작업을 분배하게 된다.

```
Viterbi decoding ()
{
 for every observation feature
   for every active GMM  // 멀티 코어
     evaluate GMM score; // SIMD
   for every active HMM  // 멀티 코어
     evaluate HMM score; // SIMD
       …
}
```

(그림 4) 비터비 디코딩 병렬처리 예

〈표 3〉 SIMD 명령어를 사용한 병렬처리 실험 환경

| | | |
|---|---|---|
| H/W | CPU | Xeon 5550 |
| | CPU clock | 2.67GHz |
| | Memory clock | 1333MHz |
| 음향 모델 | 특징벡터 차수 | 39 |
| | Gaussian mixture 개수 | 80 |
| | 음향모델 unit 개수 | 11392 |

〈표 4〉 SIMD 명령어를 사용한 병렬처리 실험 결과

| | 평가 모듈 (xRT) | |
|---|---|---|
| 명령어 | GMM | GMM+ HMM |
| SISD | 0.851 | 0.898 |
| SIMD | 0.389 | 0.407 |

(그림 4)는 비터비 디코딩 알고리즘을 SIMD 명령어와 멀티 코어를 사용해 구현하는 의사 코드이다. 매 시간 프레임마다 4개의 GMM(Gaussian Mixture Model) 기반 확률 분포와 3개의 확률 상태로 구성된 1개의 HMM(Hidden Markov Model)을 SIMD 명령어를 사용해 동시에 계산하고 이러한 계산의 대상이 되는 모든 활성화된 GMM과 HMM을 멀티 코어로 분배하게 된다.

<표 3>과 <표 4>는 다양한 병렬처리 중 SIMD 명령어를 사용한 병렬처리에 따른 음성인식 속도 개선을 보여 주는 실험 환경 및 결과이다. 인식 속도는 RTF(real time factor)로 측정하였다.

모든 GMM과 HMM을 계산할 경우 <표 4>와 같이 SIMD 명령어를 사용한 경우가 그렇지 않은 경우, SISD(Single Instruction Single Data)에 비해 약 2.2배 정도 속도 개선이 있음을 알 수 있다[8].

## 3. 개발 사례

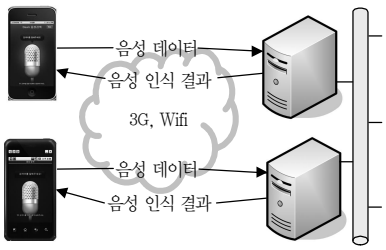
본 절에서는 ETRI에서 개발하여 포털 업체 다음에서 서비스 중인 음성검색 시스템의 개발 사례를 소개한다. 앞에서 설명한 3가지의 요소기술을 중심으로, 확률 기반의 음성인식은 언어모델, 음향모델에 따른 최대 확률 값을 가지는 단어열을 찾게 된다. 음향 및 언어 확률 모델이 결합된 단어열기반의 탐색 공간을 구성하고 이 공간 내에서 입력된 음성에 대해 최대 확률 값을 만족하는 단어열을 구하게 되는데, 최근에는 WFST(Weighted Finite State Transducer)가 가장 널리 사용된다[9].

WFST는 상태 간의 천이를 입력 심볼과 출력 심볼로 표현하는 오토마타에 가중치를 추가한 것으로 입력 심볼 열로부터 출력 심볼 열로 변환하는 과정에서 발생하는 천이 구간의 가중치를 고려한 오토마타이다. WFST는 Composition, Determinization과 Minimization 알고리즘을 사용해 음향 및 언어모델을 결합하고 최적화하여 상태와 가중치가 포함된 천이로 표현되는 탐색 공간을 구성하여 대어휘 음성인식을 효과적으로 수행할 수 있도록 한다.

음성인식 성능은 음향 및 언어 확률 모델들이 다양한 화자, 채널, 환경 및 어휘 특성 등을 얼마나 정확히 반영하도록 훈련 및 설계되었느냐에 의해 결정되는데 다양한 특성을 반영하기 위해서는 모델의 크기가 커지는 것은 불가피하다. 예를 들어, <표 5>는 음성검색 시스템에서 사용되는 확률모델과 WFST로 표현된 탐색 공간의 실제 예를 보여주는데, 다양한 화자 및 환경 특성을 반영하기 위해 수십 개의 가우시안 확률 분포 함수로 구성된 음향모델을 사용하고 많은 수의 어휘와 다양한 문장 형태를 허용하기 위해

〈표 5〉 확률모델과 WFST 구성의 실제 예

| | | |
|---|---|---|
| 음향모델 | # of Gaussian mixtures | 80 |
| | # of phone-like units | 11392 |
| 언어모델 | # of 1-gram | 1.5M |
| | # of 2-gram | 36.8M |
| | # of 3-gram | 141.9M |
| 발음모델 | average # of pronunciation | 1.6 |
| 탐색 공간 | # of states | 312M |
| | # of arcs | 719M |




(그림 5) 서버 클라이언트 기반의 분산형 음성인식

1.5M의 인식 대상 어휘와 141.9M개의 n-gram 언어모델을 사용한다. 또한, 사용자의 다양한 발음 변이를 허용하기 위해 단어당 평균 1.6개의 다중 발성을 사용하고 있다. 이 확률모델을 결합하면 WFST로 표현되는 단일 탐색 공간이 생성되는데 위의 경우에는 312M개의 상태와 719M개의 천이가 발생한다[4].

이는 과거 전통적인 음성 받아쓰기 시스템이 64K 단어를 사용하던 것에 비해 그 규모가 방대해 스마트폰과 같은 내장형 단말기나 PC 환경에서 처리하는 것은 불가능하다. 따라서, 앞에서 소개한 요소기술 가운데 하나인 병렬처리 디코딩 기술을 기반으로 (그림 5)와 같은 서버 클라이언트 방식의 분산형 음성인식을 사용해 대용량의 음성인식을 수행한다. 분산형 음성인식에서 스마트폰이나 지능형 단말기는 입력되는 신호로부터 음성 구간을 검출하고 검출된 음성 신호를 통신망을 통해 서버로 전달하는 기능을 수행하고 실제 음성인식과 관련된 작업은 대규모 음성인식 서버에서 수행하고 그 결과를 클라이언트에 전송하게 된다.

## Ⅳ. 대화형 영어교육 서비스

본 장에서는 자연어 음성인식 기술을 이용하여 대화형 영어교육 서비스를 제공하는 사례 및 개발과정에 대해 기술한다. <표 6>은 가상현실을 이용하는 대화형 언어교육에 대한 특징을 정리한 것으로[10], 이로부터 음성기술기반 대화형 영어교육 서비스의 특징을 유추할 수 있다.

〈표 6〉 가상현실을 이용한 대화형 언어교육의 장단점

| | |
|---|---|
| 장점 | - 학습 언어의 문화체험 가능<br>- 강사 부족 문제 절감 가능<br>- 흥미유발 및 장기간 기억 가능<br>- 실수에 대한 부담 감소로 자유로운 학습 가능<br>- 시간/공간적 제약에서 자유로움 |
| 단점 | - 제한적인 body language<br>- 기술적 문제(시스템 결함, 반응 속도 개선 등)<br>- 교육의 효용성 증명 부족 |

### 1. 대화형 영어교육 서비스 동향

음성인식 기술 성능의 향상과 함께 1990년대 말부터 음성인식 기술기반 언어교육 기술이 연구되어 왔으며, 자연어처리 기술 등 다른 기술과 연동 및 확장되었다. 또한, 2000년대 초부터 음성인식 기술기반 언어교육 서비스용 실제 시스템들이 개발되고 있다[11].

먼저, 음성인식 기술기반 영어 발음 평가에 대한 연구는 Fluency system(1998), ISLE(Interactive Spoken Language Education)의 CAPT(Computer-Aided Pronunciation Training)(1998~2000), Web-Grader(2000), PLASER(Pronunciation Learning via Automatic Speech Recognition)(2000~ 2002), Kyoto University의 CAPT(HUGO)(2004), EURO-NOUNCE(2007), CUHK(Chinese University of

Hong Kong)의 CAPT(2009) 등이 있다. 또한, TOTALe, Pronunciation Power, PhonicsTutor, Eyespeak, GnB online Program, iWinner Application, EduSpeak, NativeAccent, ATR CALL, EnglishCentral, TELL ME MORE 등과 같은 제품들이 출시되어 있다. 뿐만 아니라, 음성인식 기술기반 말하기능력 평가 서비스로는, ETS의 SpeechRater, Versant(PhonePass) 등이 있다.

다음으로, 음성인식, 자연어처리, 가상현실기술 등 다양한 기술들이 결합되는 대화형 영어교육에 대한 연구는, SPELL(Spoken Electronic Language Learning)(1993), MILT(Military Language Tutor)(1992~1998), PROMISE(PROjekt Mediengestütztes Interaktives Sprachelnlernen Englisch)(1994), VILTS(Voice Interactive Language Training System)(1996), IWiLL(Intelligent Web-based Interactive Language Learning)(2000~), DISCO(Development and Integration of Speech technology into COurseware for language learning)(2008~2010) 등이 있다. 또한, EduSpeak, Talklish New York Story, Rosetta stone ReFLEX, TraciTalk, 한빛소프트 오디션 잉글리시, TELL ME MORE 등과 같은 제품들이 출시되어 있다.

## 2. 대화형 영어교육을 위한 요소기술

### 가. 영어 발음 평가 기술

영어 발음 평가 기술은, 일반적으로 영어 음성인식, 영어 발음 평가, 피드백 제공 등 세 과정으로 구성되며, <표 7>은 영어 발음 평가를 위한 주요 기술에 대하여 정리한 것이다.

영어 음성인식 과정에서는 주어진 발성문장과 이에 해당하는 음성 데이터에 대하여 음소별 구간 정렬을 수행하며, 이를 통하여 발성문장의 각 단어 및 음소에 대한 시간 정보와 확률 정보를 획득한다. 또한, 일반적으로 학습자가 주어진 발성문장을 발화한다는 가정이 전제되며, 임의의 발성문장에 대한 발음 평가를 위하여 일반적인 음성인식 과정이 선행되기도 한다. 그러므로 영어 음성인식 과정은, 학습자의 모국어 특성을 충분히 반영함으로써 음성인식시스템의 음소 구간 정렬 성능을 향상시키는 기술들이 요구된다.

〈표 7〉 영어 발음 평가를 위한 요소기술

| | |
|---|---|
| 영어 음성 인식 | • 학습용 영어 단어/문장 목록 구축<br>• 학습자의 음성 데이터 구축<br>• 학습자의 모국어 특성 분석<br>• 학습자의 영어 발음 특성 분석<br>• 학습자의 모국어 특성이 반영된 음향 및 언어모델 생성<br>• 영어 음성인식 기반 음소별 구간 정렬 기술 |
| 영어 발음 평가 | • 발음 평가에 적절한 정보 분석<br>– 각 음소별 확률값<br>– 반음소모델에 대비 음소별 확률값<br>– 음소 구간 길이/에너지<br>– 묵음 및 휴지 구간 길이/횟수<br>– 포만트/강세/억양/운율<br>• 학습자의 발음 평가 기술<br>– 표준 발음열 및 음소열 사이의 유사도<br>– 음소단위의 우도비(likelihood ratio)<br>– SVM(Support Vector Machine), CART(Classification and Regression Trees), LDA(Linear Discriminant Analysis) 등의 통계적 분류 방법론을 이용한 유사도 측정 |
| 피드백 제공 | • 음소/단어/문장마다 평가 결과 제공<br>– 표준 음소열<br>– 원어민과 학습자와의 발음 비교<br>• 발성문장 내의 단어별 정보 제공<br>– 단어별 발음 표기<br>– 원어민의 발성 영상<br>– 원어민 음성 또는 합성 음성의 오디오<br>• 오류 음소에 대한 발음 정보 제공<br>– 발성 방식 설명(텍스트/그림/동영상) |

영어 발음 평가 과정에서는 영어 음성인식 과정으로부터 획득된 발성문장의 단어 및 인식된 음소열에 대한 시간 정보, 확률 정보를 사용하고, 강세, 억양, 운율, 길이 등 부가적인 정보를 이용함으로써 학습자의 발성음성에 대한 발음 평가를 수행한다. 그러므로

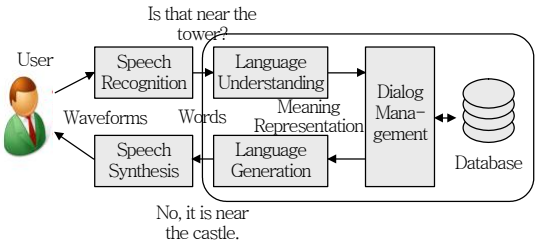
영어 발음 평가 과정은 인식된 음소열의 발음 평가에 효과적인 정보 획득, 원어민 화자의 발성음성과 학습자의 발성음성 사이의 발음 유사도를 측정하기 위한 척도(GOP: Goodness of Pronunciation) 등이 요구된다.

피드백 제공 과정에서는 학습자의 이해 및 학습을 위하여 영어 발음 평가 결과 및 부가적인 정보를 출력한다. 피드백 제공 과정은 발음 평가 결과를 효과적으로 제시하는 것과 더불어 영어 발음 학습을 위하여 제공되어야 할 정보, 오류 음소에 대한 설명, 학습자에게 친밀한 그래픽 인터페이스 등에 대한 기술 등이 요구된다.

### 나. 대화 처리 기술

외국어 학습에서 외국인과 직접 대화를 하거나 또는 역할 놀이 등의 학습으로 반복적으로 외국어를 사용하는 것이 매우 효과적이다. 이러한 외국인과의 대화 상대나 역할 놀이 상대로 음성 대화 시스템을 외국어 교육에 활용하고자 하는 시도가 있어왔다[12],[13],[14]. 대화 생성 기술의 완성도 부족으로 음성 대화 시스템이 가지는 도메인 제한적인 문제도 외국어 교육에서는 도메인에 대한 몰입 교육으로 효과가 높을 것으로 기대하고 있다[12]. 대표적인 예로, 물건을 구입하는 어휘를 학습한 후, 벼룩시장 판매원인 기계를 대상으로 학습 어휘를 반복적으로 활용하는 DEAL 시스템을 들 수 있다[13]. 대화 시나리오와 규칙, 대화생성이 제한적이지만 반복적인 언어 활용에 도움이 될 것이다.

(그림 6)은 음성 대화 시스템의 구성도이다. 음성 대화 시스템은 크게 음성인식, 대화 처리와 음성합성으로 이루어져 있으며, 대화 시스템에 해당하는 대화 처리 부분은 다시 언어이해, 대화관리와 언어생성으로 구성되어 있다. 언어이해 모듈은 사용자의 발화를




(그림 6) 음성 대화 시스템 구성도

언어처리를 통하여 분석하여 사용자의 의도를 표현하는 의미표현을 생성한다. 대화관리 모듈은 사용자와의 대화 흐름과 의도에 따른 최선의 대화 전략을 계산하여 다음 시스템 대화에 해당하는 의미표현을 생성한다. 언어생성 모듈은 시스템이 발화할 문장을 입력된 의미표현으로부터 생성한다.

대화 시스템에서 인간과 기계와의 대화가 진행하도록 하는 대화관리 방법이 핵심 요소이며, 대화관리 방법 기술에 따라 언어이해와 언어생성 방법론이 정해진다. 대화관리 방법은 대화 시스템이 적용되는 특정 도메인의 목적에 알맞은 대화 전략을 전문가에 의하여 직접 구축하는 지식기반 방법론(knowledge-based approach)과 특정 도메인의 코퍼스 데이터로부터 학습 또는 예제화하여 대화전략을 구축하는 데이터 주도 방법론(data-driven approach)으로 구분할 수 있다. 지식기반 방법론으로는 finite-state, form-filling, information-state-update와 plan-based approach가 있으며[15], 데이터 주도 방법론으로는 POMDP(Partially Observable Markov Decision Process)[16]와 EBDM(Example-Based Dialog Modeling)[17]이 있다.

일반적으로 대화관리 방법론들은 버스정보 안내, 비행기 예약 정보 안내 및 여행 정보 안내 등과 같은 정보 서비스를 위한 대화형 인터페이스에 맞도록 설계되었다. 이러한 정보서비스 대화 시스템은 목적을 빨리 달성하는 대화 흐름으로만 대화 전략이 구축 또

는 학습되어 있어서, 사용자의 발화에 대하여 같은 대답과 같은 대화 흐름으로만 일관하게 된다. 또한, 기계적인 딱딱한 대화생성은 대화 시스템을 외국어 교육 목적으로 사용하기에는 매우 부적합하다.

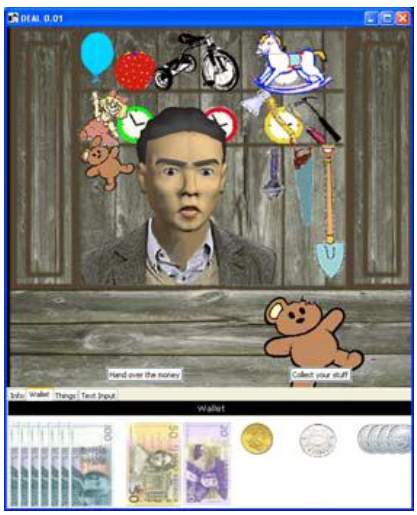
ETRI에서는 대화 시스템을 영어교육과 같은 외국어 교육에 활용하기 위해, 실제 대화의 다양성과 유창성을 표현하도록 하여 사람과 같은 자연스러운 대화가 가능한 동적 대화 그래프기반 대화 시스템을 개발하고 있다. 동적 대화 그래프 기반 방법론은 특정 도메인에서의 실제 대화들을 수집하여, 시스템과 학습자에 해당하는 대화 문장들을 대화 이력과 대화 의미 표현에 따라 같거나 다른 정점(vertex)들로 학습하고, 대화 코퍼스에 연속적으로 나타나는 학습된 대화 정점들 간을 방향성 고리로 연결하여 각 도메인에 적합한 대화 그래프를 학습한다. 학습한 대화 그래프를 이용하여 대화 시스템에서는 학습자 발화에 해당하는 정점을 대화 그래프에서 찾아낸 후 사용자 정점과 연결된 적절한 시스템 발화를 선택하여 다음 대화를 진행한다. 임의의 순간에 시스템 발화와 연결된 학습자 발화 정점들은 학습자에게 다음 순간에 자신이 할 수 있는 대화들을 제시할 수 있다. 동적 대화 그래프 기반 대화 처리 방법은 기존 정보 서비스 대화 시스템에서 항상 최적의 대화 전략만을 제시하는 것과는 달리, 학습자 수준에 따라 또는 교육 방법에 따라 대화 흐름을 그래프를 통하여 직접 통제할 수 있는 특징을 가진다. 또한, 대화 그래프를 도메인 코퍼스로부터 직접 학습을 하므로, 도메인 확장이 쉬워 다양한 영어 학습 주제를 구현하는데 효과적일 것으로 기대된다.

## 3. 개발 사례

본 절에서는 자연어 음성인식 기술을 이용하여 대화형 언어학습을 수행하는 시스템의 외국 사례를 소개하고, 자연어 음성인식 기술과 동적 대화 그래프 기술을 이용하여 대화형 영어교육이 가능하도록 ETRI에서 개발한 시스템을 소개한다.

스웨덴 KTH(Kungliga Tekniska högskolan)에서는 제2언어(영어) 학습자를 위한 역할 놀이 대화 시스템 DEAL을 개발하였다[13]. DEAL 시스템은 finite state 네트워크 기반 대화 관리 기법을 이용하여 개발되었으며, 특정한 사건이 발생하면 정해진 상태가 전이하는 방식으로 대화를 관리한다. DEAL은 먼저 벼룩시장 도메인에 맞는 기본 어휘 및 표현을 학습하고 학습한 어휘 및 표현을 이용하여 가까운 벼룩시장에 가서 주어진 물품을 구입하는 역할을 수행하도록 구현되어 있다. DEAL은 (그림 7)과 같은 GUI 인터페이스를 통하여 학습자와 대화를 진행하도록 한다. 학습자에게 주어진 돈이 부족하므로 흥정을 해서 가격을 깎아서 구입하도록 하여 학습자의 흥미를 유도하도록 설계되어 있다.

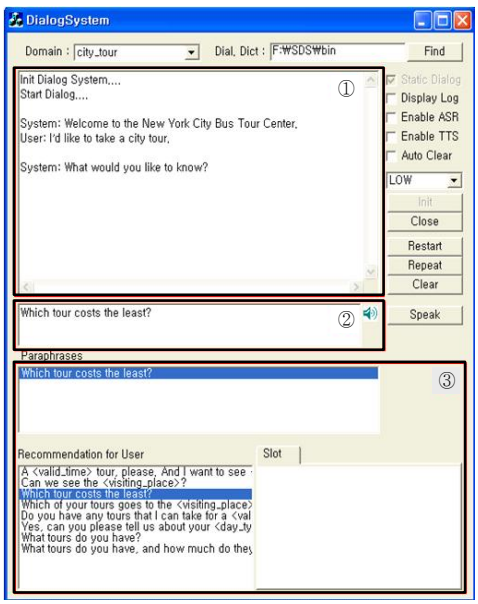
MIT에서는 몰입식 제2언어 학습을 위하여, 가족관계에 몰입해서 중국어를 학습하도록 하는 Family ISLAND를 개발하였다[12]. Family ISLAND는 가



(그림 7) DEAL 인터페이스

족 관계도를 학습자에게 제시하여 4단계의 수준별로 중국어를 학습하도록 한다. 1, 2 단계에서는 가족 관계도를 이용한 간단한 단어 학습을 하고, 3단계에서는 가족관계에 대한 시스템 질문에 대답하는 학습을 수행하고, 4단계에서는 사용자가 질문을 하고 시스템이 대답하는 학습을 진행하도록 구성되어 있다. MIT에서는 음성인식 오류나 대화처리 오류가 학습자의 교육에 치명적일 수 있다는 관점에서 오류가 적은 아주 제한적 도메인과 시나리오로 오류를 최소화한 것이 특징이다. 또한 웹 기반으로 대화 시스템과 대화를 통하여 다양한 게임을 진행하여 제2언어를 습득할 수 있으며, 필요에 따라 학습자에게 고성능 자동통역 시스템의 결과를 제공하는 교육용 대화 시스템에 대한 연구를 진행 중에 있다[14].

ETRI에서는 대화형 영어 교육시스템을 외국여행에서 쉽게 접할 수 있는 도메인들을 대상으로 개발하고 있다. 도메인 선정은 대학생을 대상으로 설문조사 등을 통하여 교육적 필요성과 효과 및 흥미를 함께 고려하여 10개 도메인을 선정하였다, 필요성의 측면에서는 외국 여행 등에서 자주 접하는 도메인으로, 효과의 측면에서는 해당 도메인이 다양한 상황 속에서 교육자로 하여금 다양한 발화를 요구하여 교육적 효과가 큰 도메인으로 고려하였다. 현재 ETRI에서는 도시관광, 입국수속 도메인을 대상으로 음성대화형 영어교육 프로토타입 시스템을 개발하였다. (그림 8)은 ETRI에서 개발한 도시관광 도메인에 대한 영어 교육용 음성 대화 시스템의 실행 예를 나타낸다. (그림 8)에서 ①은 메인 윈도로서 시스템과 학습자 간의 대화 내용이 표시된다. ②는 학습자 발화에 대한 음성인식 결과 또는 텍스트 입력이 표시된다. ③은 영어에 익숙하지 못한 학습자를 위해서 현재의 대화시점을 기준으로 해서 학습자가 발화할 수 있는 예문들을 제시하여 준다.



(그림 8) 영어 교육용 음성 대화 시스템 실행 예

## V. 결론

음성 대화 서비스에서의 자연어 음성인식 기술의 역할을 소개하고, 사용자 경험 및 로그데이터를 바탕으로 인식 성능을 지속적으로 개선하는 선순환 구조를 살펴보았다, 또한 음성검색 서비스 및 대화형 영어교육 서비스를 통해 음성 대화 서비스의 필요성을 살펴보았다. 모바일 기술 및 클라우드 시스템의 폭발적인 성장으로 음성 대화 시스템을 이용한 자연스러운 인터페이스 및 대화형 점보검색 서비스의 요구가 증가하고 있는 상황에서, 자연어 음성인식 기술을 발전시키는 선순환 구조를 구축하고 음성 대화 서비스의 보편화를 위한 방향을 제시하고자 하였다. 개발 사례로 든 음성검색 서비스 및 대화형 영어교육 서비스를 통해 무제한급 자연어 음성인식과 음성 대화 시스템의 가능성을 엿보았고, 최종적으로는 사용자 경험과 데이터를 기반으로 사람이 수행하는 수준으로 발전을 기대한다.

**● 용 어 해 설 ●**

자연어 음성인식 기술: 화자, 환경, 어휘에 무관하게 누구나 자연스럽게 발성한 음성을 인식하는 기술

음성 대화 시스템: 자연어 음성인식을 기반으로 사용자의 발성을 이해하여 대화를 통해 사용자가 원하는 정보를 제공하거나 원하는 기능을 수행하는 시스템

## 약어 정리

| | |
|---|---|
| CAPT | Computer-Aided Pronunciation Training |
| CART | Classification and Regression Trees |
| CUHK | Chinese University of Hong Kong |
| DISCO | Development and Integration of Speech technology into COurseware for language learning |
| EBDM | Example-Based Dialog Modeling |
| GMM | Gaussian Mixture Model |
| GOP | Goodness of Pronunciation |
| GPU | Graphics Processing Unit |
| HMM | Hidden Markov Model |
| ISLE | Interactive Spoken Language Education |
| IWiLL | Intelligent Web-based Interactive Language Learning |
| KTH | Kungliga Tekniska högskolan |
| LDA | Linear Discriminant Analysis |
| MILT | Military Language Tutor |
| MPI | Message Passing Interface |
| OpenMP | Open Multi-Processing |
| PLASER | Pronunciation Learning via Automatic Speech Recognition |
| POMDP | Partially Observable Markov Decision Process |
| PROMISE | PROjekt Mediengestütztes Interaktives Sprachelnlernen Englisch |
| RTF | real time factor |
| SaaS | Speech As A Speech |
| SIMD | Single Instruction Multiple Data |
| SISD | Single Instruction Single Data |
| SPELL | Spoken Electronic Language Learning |
| SSE | Streaming SIMD Extension |
| SVM | Support Vector Machine |
| VILTS | Voice Interactive Language Training System |
| WFST | Weighted Finite State Transducer |

## 참 고 문 헌


[1] Elise Ackerman and Erico Guizzo, "5 Technologies That will Shape the Web," *IEEE Spectr.*, June 2011, pp. 33-37.

[2] Jeff Ma and Spyros Matsoukas, "Unsupervised Training on a Large Amount of Arabic Broadcast News Data," *Proc. ICASSP*, 2007, vol. II, pp. 349-352.

[3] Thorsten Brants et al., "Large Language Models in Machine Translation," *Proc. Joint Conf. Empirical Methods Nat. Language Proc. Comput. Nat. Language Learning (EMNLP-CoNLL)*, 2007, pp. 858-867.

[4] Ahmad Emami, Kishore Papineni, and Jeffrey Sorensen, "Large-Scale Distributed Language Modeling," *Proc. ICASSP*, 2007, vol. IV, pp. 37-40

[5] Jeffrey Dean and Sanjay Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters," *6th Symp. Operating Syst. Design Implementation (OSDI-04)*, San Francisco, CA, USA, 2004.

[6] S. Kanthac, K. Schutz, and H. Ney, "Using SIMD Instructions for Fast Likelihood Calculation in LVCSR," *Proc. ICASSP*, 2000, vol. 3, pp. 1531-1534.

[7] K. You et al., "Parallel Scalability in Speech Recognition: Inference Engine in Large Vocabulary Continuous Speech Recognition," *IEEE Signal Proc. Mag.*, no. 6, Nov. 2009, pp. 124-135.

[8] 정훈 외, "고속 음성인식을 위한 병렬 처리," 대한음성학회 2011년 춘계학술대회, 2011.

[9] Mehryar Mohri, Fernando Pereira, and Michael Riley, "Weighted Finite-State Transducers in Speech Recognition," *Comput. Speech Language*, vol. 16, no. 1, 2002, pp. 69-88.

[10] Stefanie Hundsberger, "Foreign language learning in Second Life and the Implications for Resource Provision in Academic Libraries," Arcadia Fellowship Programme, Cambridge University Library, June 2009.

[11] Maxine Eskenazi, "An Overview of Spoken Language Technology for Education," *Speech*

*Commun.*, vol. 51, Oct. 2009, pp. 832-844.

[12] Ian McGraw and Stephanie Seneff, "Immersive Second Language Acquisition in Narrow Domains: A Prototype ISLAND Dialogue System," *Proc. SlGSLaTe Workshop,* Farmington, PA, Oct. 2007.

[13] Preben Wik, Anna Hjalmarson, and Jenny Brusk, "DEAL a Serious Game for CALL Practicing Conversational Skills in The Trade Domain," *Proc. SlGSLaTe Workshop,* Farmington, PA, Oct. 2007

[14] Stephanie Seneff, "Web-Based Dialogue and Translation Games for Spoken Language Learning," *Proc. SlGSLaTe Workshop,* Farmington, PA, Oct. 2007

[15] Dan Bohus and Alexander I. Rudnicky, "The RavenClaw Dialog Management Framework: Architecture and Systems," *Comput. Speech Language,* vol. 23, no. 3, 2009, pp. 332-361.

[16] Steve Young et al., "The Hidden Information State Model: A Practical Framework for POMDP-based Spoken Dialogue Management," *Comput. Speech Language*, vol. 24, no. 2, 2010, pp. 150-174.

[17] Cheongjae Lee et al., "Hybrid Approach to Robust Dialog Management Using Agenda and Dialog Examples," *Comput. Speech Language*, vol. 24, no. 4, 2010, pp. 609-631.