대화음성 인터페이스 기술 및 응용 서비스 개발 동향


정 호 영
ETRI 자동통역
인공지능연구센터


윤 승
ETRI 자동통역
인공지능연구센터


전 형 배
ETRI 자동통역
인공지능연구센터


이 윤 근
ETRI 자동통역
인공지능연구센터


박 기 영
ETRI 자동통역
인공지능연구센터


박 전 규
ETRI 자동통역
인공지능연구센터


김 운
ETRI 자동통역
인공지능연구센터

Ⅰ. 서 론

언어는 인간에게 있어서 가장 중요한 정보전달 및 의사소통 수단이다. 수십년 전부터 인간의 언어활동을 컴퓨터로 모사하려는 시도가 이루어져 왔으며 최근 들어 산업 사회가 지식 서비스 사회로 접어들면서 이에 대한 관심은 더욱 높아지고 있다. 이러한 분야의 기술을 음성언어처리기술 이라고 부르며 대표적인 예로는 음성인식/합성 기술, 대화처리 기술, 자동번역 기술 등이 있다.

음성언어처리 기술은 인간의 언어를 다루는 기술이므로 HCI(Human Computer Interaction)의 핵심 기술이며 지식 및 정보서비스의 기반 기술이며 나아가 문화산업의 기반이 되는 기술이니만큼 이미 선진국에서는 그 중요성에 대해 깊이 인지하고 있으며 핵심기술 개발에 많은 노력을 기울이고 있다. 세계적인 IT 기업인 애플, 구글, 마이크로소프트, IBM 등이 모두 음성언어처리 기술 확보에

많은 노력을 기울이고 있으며 핵심 기술을 이용한 다양한 서비스를 선보이고 있다. 국내에서도 국가출연연구소를 중심으로 핵심기술에 대한 연구가 지속적으로 이루어지고 있으며 일부 기술 분야에서는 선진국 기술 수준에 비하여 뒤떨어지지 않는 경쟁력을 보유하고 있다.

본 논문에서는 앞에서 언급한 음성언어처리기술에 대한 최근 동향을 살펴보고 국내에서 개발된 대표적인 음성언어 기반 서비스에 대하여 소개하고자 한다.

Ⅱ. 음성인식 기술의 최근 동향

모바일 정보서비스의 성장으로 음성인식을 기본적인 인터페이스로 생각하는 사용자들이 늘어가고 있다. 편리한 인터페이스에 기반한 정보 검색 서비스를 위해 음성인식을 채택하는 경우가 점차 늘어가고 있으며, 사용자 로그 분석에 따라 음성인식 성능도 지속적으로 향상되어 가고 있다. 본 장에서는 최근 관심을 받고 있는 빅데이터에 기반한 음성인식 기술 동향에 대해 살펴보고자 한다.

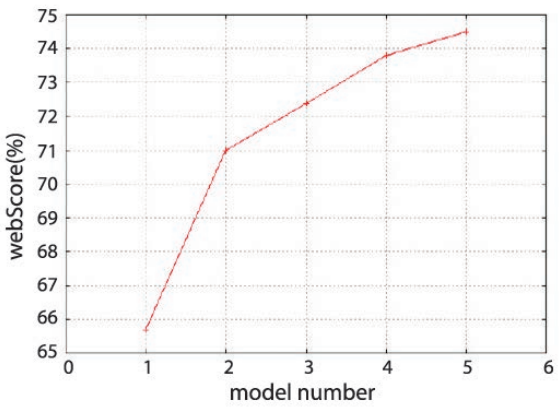
음성인식 기술은 명령어 인식의 수준에서 시작하여 낭독체 연속어 인식, 대어휘 연속어 인식을 거쳐 무제한급의 자연어 음성인식의 단계로 발전하고 있다. 음성인식 기술의 패러다임은 스마트폰의 활성화와 더불어 구글에서 공개한 음성검색 서비스를 기준으로 나누어 볼 수 있다. 2007년 이후 다양한 정보에 편리하게 접근하기 위한 사용자 인터페이스의 필요성과 함께 모바일 환경에서의 빠른 정보검색이라는 요구가 맞물리면서 음성검색 서비스가 개발되어 왔다. 이와 같은 서비스를 위해서는 다양한 환경적 특성 및 화자적 특성을 아우르는 음향 모델링 방법론이 필요하며 사용자 발화 분석에 따른 언어 모델링 전략을 기반으로 자연어 음성을 인식할 수 있는 수준의 기술이 요구된다. 2007년 이전이 음성인식 성능 개선을 위한 다양한 통계적 방법론을 시도한 시기라면, 2007년 이후는 다양한 데이터들을 기반으로 개발된 통계적 방법론의 한계를 해결하려는 시기라고 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 방대한 데이터를 이용하여 통계적 모델을 구축하고 어휘 탐색 공간을 결정하여 음성인식을 수행하는 방법론이 음성인식 기술 연구의 한 축을 이루고 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.

빅데이터에 기반하여 음성인식 성능 문제를 해결하려는 시도는 구글을 중심으로 이루어지고 있다. 구글은 검색 서비스를 기반으로 하여 방대한 규모의 음성 데이터 및 텍스트 코퍼스를 수집하고 있으며, 이를 이용하여 음향모델 및 언어모델을 구축하는 방법론을 채택하고 있다. 최근 구글의 발표에 따르면 매일 쌓이는 음성 데이터의 양은 한사람이 5~10년 동안 말하는 양에 해당하며, 텍스트 코퍼스의 양은 음성데이터를 넘어서고 있다. 이런 데이터를 이용하여 통계적인 모델링을 수행하는 방법은 기존에 발표된 기술과 크게 다르지 않으며, 기존에 제한된 데이터로 성능 개선을 이루지 못했던 것이 데이터의 확보로 성능 개선을 이루었다는데 의미를 얻을 수 있다. 음향 모델을 구축하기 위한 방법은 HMM(Hidden Markov Model)에 기초로 boosted MMI(Maximum Mutual Information) 기법 등의 변별학습 기술을 도입하는 것이고, 언어 모델을 구축하는 방법은 n-gram 기술을 그대로 이용하고 있다. 한가지 주목할 만한 점은 방대한 데이터로 인해 음성데이터에 대한 정확한 전사 데이터가 제공되지 않고 텍스트 코퍼스에 대한 정확한 세그멘테이션이 제공되지 않는다는 것이다. 미전사 데이터 문제를 해결하기 위해서는 supervised 훈련을 넘어선 unsupervised 훈련 방법이 시도되고 있다. 초기 모델은 적은 분량의 전사 데이터를 가진 음성 데이터를 이용하여 supervised 훈련을 통해 생성하고, 초기 모델을 이용하여 방대한 양의 미전사 데이터를 자동 전사하면서 다양한 신뢰도 척도를 이용하여 신뢰도를

계산한 후 높은 신뢰도를 가진 데이터를 훈련에 반영하는 방법을 이용하고 있다. 정확하게는 semi-supervised 훈련이라고 말할 수 있으며, 구글에서 로그 데이터를 처리하는 노력의 대부분이 여기에 집중되고 있다. 구글은 음성인식 엔진에 사용되는 모든 음향모델을 이 과정을 통해 4주마다 재학습하고 있으며, 최근의 사용자 사용 특성을 반영하는 미전사 음성데이터를 주기적으로 반영하고 있다. 데이터가 축적됨에 따라 성능 개선 정도는 점차 줄어들고 있으나 꾸준한 성능 개선을 보이고 있다. 〈그림 1〉은 구글에서 개발한 음성 검색 서비스의 음향모델 개발 단계를 보여주는 것으로, 2000 시간의 전사를 가진 음성데이터 모델을 기반으로 모델 3부터 5,000 시간 이상의 미전사 음성데이터를 추가하면서 꾸준한 성능 향상을 보이고 있다[1].

텍스트 코퍼스를 이용하여 언어모델을 구축하는 것은 방대한 양의 데이터를 처리하는 방법에 달려있다. 구글의 경우 수집한 텍스트 코퍼스로부터 2,300억개의 단어를 얻을 수 있었는데, 이를 기반으로 효과적으로 n-gram 기반의 언어모델을 구축하는 방법론을 도출하였다. 초기의 언어모델은 백만개의 어휘를 선택하고 이를 기반으로 3-gram까지 표현하는 1,500만개의 n-gram을 구성하였다. 이렇게 구축된 언어모델은 음성인식 과정의 첫 번째 인식에 적용되고, 첫 번째 인식과정에서 얻은 후보들을 대상으로 127억개의 n-gram을 가진 5-gram 으로 리스코어링하는 방법을 적용하였다. 방대한 텍스트 코퍼스를 이용하여 얻을 수 있는 어휘 가운데 사용 빈도가 높은 어휘를 선정하고 이를 기반으로 n-gram 개수를 다양한 단위로 확장하면서 적용하는 것이다. 결국 언어모델의 방법은 대상 어휘의 선정과 확장 가능한 n-gram의 구성 및 미출현 n-gram에 대한 스무딩 기법에 따라 성능 개선을 얻을 수 있음을 알 수 있다.

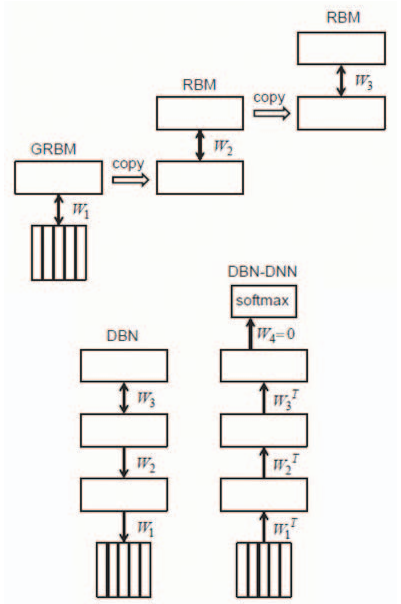


〈그림 1〉 음향모델에 따른 구글 음성검색 서비스의 성능

한국어의 경우는 음성인식 과정의 효율을 위해 단어가 아닌 형태소 단위의 어휘를 선정하게 되는데, 이 경우 어휘 선정의 문제와 함께 형태소 단위의 분할도 성능에 영향을 끼치게 된다. 방대한 양의 텍스트 코퍼스를 자동 형태소 분할하는 경우 텍스트 자체의 오류와 분할 오류 등이 더해져 의미없는 어휘가 생성되는 경우가 자주 발생한다. 따라서 한국어의 경우는 형태소 분할, 어휘 선정 및 n-gram 의 확장 구성에 대한 고려가 이루어져야 언어모델의 성능을 개선할 수 있다.

HMM 에 기반한 전통적인 음향모델의 성능을 개선하기 위한 DNN(Deep Neural Network) 기술이 최근에 주목받고 있다. 기존에 시도되었던 HMM과 신경회로망 기술의 결합으로 다중 layer 기반의 신경회로망 학습에 한계가 있어 음성인식 성능 개선에 큰 기여를 하지 못하다가, 최근 들어 DNN 기술의 개발로 인해 음성인식 성능을 대폭 개선하는 효과를 보이고 있다. DNN을 HMM 과 결합하는 방법은 HMM의 각 state 확률 분포를 모델링하는 방법으로 널리 사용되던 GMM(Gaussian Mixture Model)을 DNN으로 대체하는 것으로, 특징으로부터 state 확률을 계산하는데 DNN을 적용하는 것이다. DNN의 입력은 10여 프레임의 음성특징 벡터가 되고 출력은 HMM에서 단위로 사용하는 모든 트라이폰의 각 state가 된다. 10여개 프레임으로부터 얻은 특징벡터열을 입력하여 학습된 DNN의 layer의 모델 파라미터에 따라 최종 출력에서 트라이폰 각 state의 확률값을 얻을 수 있고, 이를 이용하여 등록된 어휘 단위의 인식결과를 탐색하게 된다. 아래에 있는 〈그림 2〉는 DNN을 이용하여 특징벡터로부터 HMM state를 학습하는 과정을 나타낸다.

음성특징을 입력으로 하여 RBM(Restricted Boltzmann Machine)으로 layer 구성하면서 layer 별



〈그림 2〉 DNN을 이용한 음향모델 학습 과정

로 학습을 수행하고 학습된 layer를 모두 모아 DBN(Deep Belief Network)을 학습하는 것을 pre-train 이라 하며, 이렇게 구성된 DBN-DNN 모델에 최종 트라이폰 state별 전사정보를 주어 최종 layer를 역전파알고리즘을 이용하여 학습하면 최종 DNN 모델을 얻게 된다. DNN-HMM 구조는 기존의 GMM-HMM 기법에서 성능 향상을 위해 적용했던 다양한 특징 정규화 기법 및 변별학습 방법을 layer 학습을 통해 표현할 수 있는 장점을 가지며, 이로 인해 기존에 해결하지 못했던 문제들을 layer 학습내에서 처리할 수 있다. 실제로 구글에서는 DNN 기술을 적용하여 음성검색 서비스에 대해 개선된 음성인식 성능을 얻을 수 있었다. 기존의 GMM-HMM방법으로 5,870 시간 이상의 음성 데이터를 학습에 적용한 경우 16%의 음성인식 오류율을 보인 반면, DNN-HMM 기술을 적용하여 5,870 시간의 학습 데이터로 12.3%의 오류율을 확보하였다. 음성 검색이라는 도메인내에서 GMM에 기반한 방법보다 적은 데이터를 이용하여 음성인식 성능을 개선하는 효과를 보이고 있다[2].

2011년 IBM이 만든 슈퍼컴퓨터 왓슨은 제퍼디 쇼에 출연해 역사상 최고 성적을 올린 두 명의 인간과 퀴즈 대결을 벌여 우승을 하면서 세상을 놀라게 한 바가 있다.

Ⅲ. 언어처리 기술의 최근 동향

최근 들어 자동번역, Q/A 시스템, 대화시스템 등 언어처리 응용 기술들이 점점 사람들의 각광을 받고 있다. 2011년 IBM이 만든 슈퍼컴퓨터 왓슨은 제퍼디 쇼에 출연해 역사상 최고 성적을 올린 두 명의 인간과 퀴즈 대결을 벌여 우승을 하면서 세상을 놀라게 한 바가 있다. 애플은 스마트폰상에서 사람과 간단한 대화를 할 수 있는 인공지능형 음성 대화시스템 시리(Siri)를 출시하였다. 이런 기술들이 각광을 받는 주요한 이유는 언어처리 응용 기술들이 점점 더 인간중심으로 스마트해지고 있기 때문이다. 그 배경에는 언어처리 요소 기술들의 성능 고도화와 빅데이터 처리가 가능하도록 하는 풍부한 언어자원이 한 몫을 하고 있다.

1. 자동번역 기술 및 동향

가. 형태소 분석 및 구문분석 기술

형태소 분석 및 구문분석 기술은 많은 자연어 처리의 기반 기술로써 크게 규칙 기반 방법과 통계 기반 방법으로 나눌 수 있다. 이 중, 통계 기반 분석은 80년대 후반의 IBM의 통계 기반 번역의 대두, 90년대 초반의 펜실베니아 대학의 Penn TreeBank 구축 등으로 큰 각광을 받기 시작했다.

초창기의 통계기반 방식은 입력과 출력의 joint생성 과정을 함께 모델링하는 생성모델(Generative Model) 기반 방법에 국한되었다. 그러나 생성 모델은 출력 구조에서 입력문으로의 생성 과정을 엄밀하게 정의해야 하기 때문에 다양한 자질을 임의로 조합하기에 불편함이 많았다. 그 대안으로, 기계학습의 분류 접근법 (Discriminative Approach)은 최대 엔트로피 모델 (Maximum Entropy Model)과 2000년도 초반 CRF(Conditional Random Field)의 시작으로 현재까지 언어처리 연구의 주된 축을 이루고 있다.

분류 접근법은 다양한 자질을 임의로 도입하여 조합할 수 있다는 점, 출력 과정을 직접 모델링하기 때문에 상대적으로 높은 성능을 낼 수 있는 점 등의 강점을 지닌다. 초창기 대표적인 분류 접근법으로는 Logistic Regression과 SVM등을 들 수 있다. 이들은 초기에 이진 분류나 Multi-class 분류 문제만 취급하는 모델이었으나, 2000년대 초반 자연언어처리의 일반적인 문제에 대응되는 구조적 분류문제를 다룰 수 있도록 재조명을 집중적으로 받아 CRF와 SVMstruct등으로 확장되었다.

최근 품사 태깅 연구는 다양한 자질을 효과적으로 결합한 CRF 및 Averaged perceptron방법에 기반을 두고 있다. 한국어 형태소 분석에서도 음절기반 CRF기반 방법이 ETRI와 강원대 및 성신여대 등에서 연구되었다[3]. 또한, 순차열 태깅으로 입력문 전체를 동시에 태깅하지 않고, 각각의 단어나 문자를 별개의 분류 대상으로 보고 개별적인 분류를 시도하는 point-wise 접근법도 연구되었다. 한편, 래티스 기반 분류 방법은 전통적인 사전 기반 형태소 분석 방식에 대해 분류적 접근을 도입한 방법이다[4]. 이는 주어진 입력문에 대한 래티스 상에서 최적의 경로를 탐색하는 기법으로, 사전이 대규모이고 학습코퍼스가 적을 때 효과적인 것으로 알려졌다.

최근 구문 분석 연구는 의존파싱 연구가 대부분을 차지하고 있다[4]. 의존 파싱은 입력 단어들 간의 의존관계를 파악하는 과정으로 구구조 파싱보다 문제가 보다 간단하다. 의존파싱 방법은 크게 그래프 기반 방식과 전이 기반 방식의 접근법으로 나눌 수 있다.

먼저, 그래프 기반 방법은 입력 단어들을 정점으로, 단어들 간의 의존 관계 여부를 에지로 취하는 방향성 그래프를 만들고, 이에 MST (Maximum Spanning Tree)탐색함으로써 파싱을 수행하는 방법이다. 다음, 전이 기반 방법은 shift-reduce파싱 과정에서 취할 수 있는 액션 (Action)을 결정하기 위해 분류기를 도입하는 방법이다. 이들 두 방법은 각각 보다 효과적으로 분석 할 수 있는 언어들이 있는 것으로 보고되고 있으나, 다국어 전체적으로 평균을 취할 시에는 서로 엇비슷한 성능을 보이고 있는 상황이다.

빅데이터 추세에 맞추어, 구문 분석 성능 개선을 위해 웹-기반 자질을 활용하는 방식도 제안되었는데, 적용 결과, 최고 성능의 의존 파서에서도 약 7%의 구분 분석 오류 감소율을 보여주었다. 또한, 학습 말뭉치를 전혀 사용하지 않고 병렬 말뭉치만을 이용하여 한 언어의 의존 파싱 모델을 다른 언어로부터 유도해내는 방법이 구글을 중심으로 연구가 진행되고 있는 추세이다.

나. 자동번역 기술 동향

전통적인 자동번역 방법은 입력문에 대한 형태소 분석과 구문분석 등 언어적 분석을 거친 후 구문적 변환을 거쳐 목적언어의 대역문을 생성하게 되는데, 이 과정에서 정교하게 구축된 언어학적 지식(규칙이나 패턴)에 많이 의존한다. 하지만, 최근에는 통계적 자동번역(Statistical Machine Translation;SMT) 기술이 그 주를 이루고 있다. 통계적 자동번역 기술은 이중언어 말뭉치로부터 단어, 구(Phrase)단위의 통계적 분석을 통해 번역 모델과 언어모델의 파라미터를 학습하고, 학습된 모델에 근거하여 최적의 번역결과를 탐색하는 디코딩 알고리즘에 의해 번역문을 생성하는 기술이다.

통계적 자동번역 기술은 구단위 통계적 자동번역에 대한 연구가 활발히 이루고 있지만, 어순 재배열에 따른 계산 복잡도 때문에 번역 성능의 한계로 다양한 노력이 시도되고 있다.

최근 들어, 통계적 자동번역 기술은 구단위 통계적 자동번역에 대한 연구가 활발히 이루고 있지만, 어순 재배열에 따른 계산 복잡도 때문에 번역 성능의 한계로 다양한 노력이 시도되고 있다. 선형 순차 모델에서는 구문 단위 기반의 번역 모델과 N-그램 기법의 장단점을 결합하여 Long-distance 어순 재배열 문제를 해결하고자 하였다. 이 모델은 번역을 선형 순차 오퍼레이션으로 처리하는 Joint 소스 채널 확률 모델에 기반하여 번역과 어순 재배열 처리를 한다. 또한, 언어학적인 분석 기술을 다양하게 접목하여 구단위의 통계적 모델에서 언어분석 기법이 들어간 구문 단위의 통계적 모델

개발이 활발히 이루어지고 있다. 대표적인 방법론으로는 계층적 구기반 모델과 STSG(Synchronous Tree Substitution Grammar) 모델이 있다. 그 중, 계층적 구기반 모델은 이중 말뭉치에서 synchronous CFG(Context Free Grammar) 규칙을 추출하여 구단위 번역 모델에 적용한 반면, STSG 모델에서는 이중 말뭉치의 각 문장쌍을 구문트리 쌍으로 표현하여 변역 모델을 학습시키는 방법이다. 이런 구문 기반의 통계적 자동번역 방법은 최근 오픈 소스 기반의 통계적 자동번역 시스템인 MOSES에서 구문 학습 모델과 구문 디코딩 모델을 또 하나의 기능으로 지원하고 있다.

규칙/패턴 기반의 자동번역은 특정 언어, 특히 어순이 다른 언어쌍에 대한 번역 성능 향상에 유리한 반면, 통계적 자동번역은 다국어 확장에 유리하다. 하지만, 자동번역 기술은 통계기반 또는 규칙기반만으로 좋은 성과를 낼 수 없으며, 현재 자동번역의 한계는 이 두 가지 방식의 장점을 어떻게 잘 접목할 수 있는가에 달려 있다는데 공감대가 형성되어 있다.

최근 하이브리딩 방법은 통계적인 방법에 언어적 지식을 추가하는 패러다임, 규칙기반의 방법에 데이터 기반 기술을 접목하는 패러다임 등 다양하게 시도되고 있다. 그러나 복수 개의 번역엔진에서 최적의 번역결과를 선택하는데 있어서 어려움을 겪고 있으며 상용화에 있어 속도, 메모리 등 문제점을 드러내고 있다.

영어 기반의 구글, IBM, MS 등에서는 시장 규모의 우위를 바탕으로 다국어 자동번역 기술 투자 및 서비스 확대와 이를 통한 시장 확장의 선순환 체계를 구축하고 있다. 구글은 SMT 기반의 59개 국어 이상의 언어에 대한 다국어 자동번역 기술을 개발하여 다국어 처리에 대한 기술적 우위를 확보하고 있다.

국내에서는 ETRI가 한국어를 중심으로 영어, 중국어, 일본어에 대한 특허문서, 기술 문서 등과 같은 특화된 분야를 중심으로 자동번역 기술이 개발되었으나, 각각의 언어쌍 번역을 위해 별도의 번역 기술이 필요한 형태로 개발되어 다국어 확장을 지원할 수 있는 기술 개발은 이루어지고 있지 않고 있었다. ㈜씨에스엘아이에서는 한/중/일/영 문서 및 대화체 자동번역 기술을 확보하고 있으며 유럽의 Systran을 인수하여 다국어 자동번역 기술 확보에 성공했다.

자동번역 솔루션은 기존의 문서 자동번역에서 최근에 스마트폰의 보급이 활발히 이루어지면서 모바일 기반의 다양한 자동번역 응용 솔루션이 시장에 나타나고 있다. Facebook은 음성인식과 번역서비스를 제공하는 Jibbigo 앱을 만든 모바일 테크놀로지를 인수하면서 자동번역 서비스를 준비 중이며, Twitter에서는 MS Bing 자동번역을 이용하여 트위터 영어, 아랍어, 이탈리아어, 스페인어 자동번역을 개시하면서 전세계에 메시지 확산을 유도할 것으로 기대하고 있다. 또한, 구글 번역 API를 이용해서 채팅 자동번역, OCR 자동번역, 영상통화 자동번역 등 다양한 스마트폰 앱들도 출현하고 있다.

국내에서는 최근 대화체 자동번역 기술 개발을 통해 일본어, 영어, 중국어에 대한 메신저 자동번역, 음성 자동통번역 서비스가 실시되고 있다. 그리고 ETRI에서는 지식학습 기반의 언어 확장이 용이한 자동통번역 원천기술 개발 과제를 수행하고 있으며[5], 이를 통해 스페인어, 불어, 독어, 러시아 등 주요 유럽어권 언어에 대한 다국어 자동번역 기술 확보를 추진하고 있다.

2. 음성대화처리 기술 및 동향

가. 음성대화처리 기술

음성대화처리기술은 사람이 사람과 대화할 때 이루어지는 인지 및 처리 과정의 각 단계가 핵심 기술로서 작용한다. 〈그림 3〉과 같이 사용자의 발화를 인식 한 후 인식된 결과물을 컴퓨터가 이해할 수 있는 형태로 이해하는 과정을 거치고, 대화처리 단계에서 어떻게 반응해야 하는 지를 결정한 후, 그에 맞는 시스템 반응을 생성하는 것이 일반적인 처리의 흐름이다.

전통적인 음성대화 시스템의 경우는 음성합성을 통해

〈그림 3〉 음성대화처리기술 흐름도

사람의 목소리로 다시 반응을 해주는 경우가 많지만, 최근에는 음성뿐만 아니라 스마트폰처럼 화면에 결과를 표시해주거나, 로봇처럼 직접 행동으로 보여주는 형태의 반응까지 모두 시스템 응답으로 포함시킨다.

나. 음성대화처리 기술 동향

최근의 음성대화 처리 기술의 주요 동향은 도움, 추천이라는 두 가지 키워드로 요약해 볼 수 있다. 애플의 Siri가 대표적은 '도움' 형 대화처리 기술이며, 구글의 Now가 '추천' 형 대화 처리 기술이라고 할 수 있다.

'도움' 형 대화처리시스템이란 사용자가 어떤 행위를 할 때 대화처리 기술이 그 행위에 대해 대화를 통해 쉽게 할 수 있도록 보조하는 시스템을 말한다. 애플 Siri의 경우 일반적으로 채팅형 Agent로 대중에 알려졌지만, 보다 핵심 기술은 대화를 통해서 스마트폰의 기본적 기능을 처리할 수 있도록 하는 것에 있다. 즉, 전화를 걸거나 문자메시지를 작성하고, 자주 사용하는 알람 예약기능 등을 자연어를 통해서 사용할 수 있도록 보조하는 것이다. Siri의 는 대화처리의 핵심 기술의 하나로 Active Ontology를 사용한다. Active Ontology는 기존의 Ontology의 개념을 발전시킨 것으로써, 대화 도메인의 개념, 테스크 플로우, 대화 상황 등을 온톨로지 형태로 표현한 후 이 지식이 실행 타임 때 검색이나, 알림 같은 실제 서비스와 연계되어 작동하게 된다.

'추천' 형 대화처리시스템이란 사용자가 어떤 상황에 있거나 행위를 할 때 그 상황에서 가장 좋은 결과를 사용자에게 추천해주는 시스템을 말한다. 정보나 컨텐츠 검색과 대부분 연계되며, 기존의 단순 키워드 질의 방식을 자연어를 통한 대화형 검색으로 서비스할 수 있도록 하는 시스템이다. 대표적으로 구글의 Now 서비스와 삼성전자의 S-Recommendation이 있다.

구글 Now의 경우 기존의 Voice Search 기능의 키워드 질의 검색을 자연어 질의 검색 수준으로 향상 시킨 기술이라고 할 수 있다. 기존에 'How's the weather'를 구글에 검색하게 되면, 해당 단어들이 많이 나타난 문서가 검색되는 것에 비해, 음성으로 검색을 시도하게 되며, 가장 상위에 현재 날씨가 요약되어 나타난다. 구글의 Conversational Search 기능은, 이전의 질의 결과를 바탕으로 계속해서 검색을 대화 형태로 진행할 수 있도록 도와준다. 바로 전의 검색결과를 바탕으로 편리하게 부가 정보를 검색할 수 있게 한다. 구글 Now 는 단순 사용자 발화뿐만 아니라, 시간, 장소, 사용이력, 교통정보 등을 바탕으로 현재 사용자에게 가장 필요한 정보를 추천하는 것을 목표로 하고 있다.

삼성전자의 S-Recommendation 기술은 스마트 TV의 컨텐츠를 추천해주는 기술이다. 이 기술은 TV 프로그램, 동영상 등을 사용자의 자연어 발화 그리고 제스쳐 등을 통해 쉽게 찾을 수 있게끔 도와주는 멀티모달 대화형 시스템이다. 자연어 발화를 통해 컨텐츠를 추천받고 이를 선택하는 과정에서 자연어 발화(First one, Second one, ...)나 손의 제스쳐를 통해 선택하거나 다른 페이지를 검색할 수 있게 한다.

Ⅳ. 국내의 대표적 상용화 사례

1. 자동통역 서비스

자동통역 기술은 언어가 서로 다른 상대방간에 의사소통이 가능하도록 하는 기술이다. 자동통역 기술은 음성인식, 자동번역, 음성합성, 이렇게 세 가지 요소 기술이 결합되어 이루어지는데, 자동통역 대상 언어마다 각각의 핵심 요소기술을 개발해야 하므로 고난이도의 복합 기술이라고 할 수 있다. 또한 세 가지 요소 기술 외에 자연언어이해 기술 및 UI와 관련된 사용자 인터페이스 관련 기술도 자동통역에서 중요한 역할을 담당하고 있다.

한국전자통신연구원에서는 2012년 10월 영/한, 한/영 자동통역 시범 서비스(지니톡)를 처음으로 시작하였다.

가. 자동통역 서비스 개요

한국전자통신연구원에서는 2012년 10월 영/한, 한/영 자동통역 시범 서비스(지니톡)를 처음으로 시작하였다. 이후 2013년 4월 일/한, 한/일 자동통역 서비

〈그림 4〉 자동통역 앱 '지니톡' 실행 화면

스를 추가하였고, 12월에는 한/중, 중/한 자동통역으로 그 대상 언어를 확장한 바 있다. 현재 iOS와 Android OS 환경에서 서비스되고 있는데, 실행 화면은 〈그림 4〉와 같으며 다운로드 건수는 170만 건에 달한다. 국내 이용자가 약 80%를 차지하지만 전 세계 198개 국가에 이용자가 분포되어 있으며, 사용자가 1,000명 이상인 국가도 18개 국가에 이를 정도로 널리 사용되고 있다.

지니톡의 경우 한국어, 영어, 일본어 음성인식 엔진은 세 언어를 합쳐 동시에 160명이 수용 가능하도록 서비스를 실시하고 있으며 중국어 음성인식 엔진의 경우에는 20명의 동시 접속자를 지원하고 있다.[1)]

자동번역 엔진의 경우 각 언어쌍 별로 동시 접속자 16명을 수용하는 구조로 설계되어 있는데, 이는 텍스트만을 그 처리 대상으로 하므로 수행 속도가 음성인식 엔진에 비해 매우 빠르기 때문이다. 음성합성 엔진의 경우 한/영 자동통역 서비스 사용자가 많은 것을 고려해 영어의 경우 10채널, 나머지 언어는 5채널을 지원하고 있다. 현재 지니톡 서버에는 서비스 시작 이후 월평균 290만건의 자동통역 로그가 축적되고 있다. 이렇게 확보된 로그는 음성인식 및 자동번역 엔진의 성능 개선을 위해 활용된다.

나. 음성인식 엔진

지니톡에 채용된 음성인식 엔진은 HMM 기반의 음향 모델과 n-gram 기반의 언어 모델을 채택하고 있다. 그리고 디코더는 wFST(weighted Finite State Transducer) 방식으로 개발되었다.

한국어 음향 모델의 훈련에는 전체 768 시간 분량의 음성 DB가 사용되었다. 여기에 잡음 환경에서 강인하게 동작할 수 있도록 SNR 5~15 dB의 실제 환경 잡음을 무작위로 섞어 전체 1,424시간 분량의 음성 DB를 사용하여 훈련하였다. 음성 신호는 10ms마다 프레임을 이동하며 20ms 단위로 특징 추출을 실시하였고 특징 계수의 경우 MFCC 53차를 사용하였으며 이후 triphone-tying 후의 GMM(Gaussian Mixture Model) 개수는 32개까지 늘렸다. 특기할 만한 것은 자동통역을 위한 음성인식 엔진이 스마트폰에서의 대화체 음성인식이 주를 이룬다는 점에서 음향 모델 훈련에 사용한 음성 DB 중 61%를 스마트폰 채널의 대화체 데이터로 구성하여 최대한 사용 환경을 일치 시키도록 하였다는 점이다. 영어, 일본어, 중국어 음향 모델도 훈련에 사용한 DB의 구성을 제외하고는 동일한 방식을 채택하였다.

지니톡 언어모델은 대화체 자동통역에 초점을 맞추어 여행/일상 관련 대화 텍스트를 중심으로 구성하였다. 이를 위해 크게 세 가지 방법을 동원해 언어모델용 말뭉치를 수집하였다. 첫 번째로는 다양성을 최대한 확보하고 자연성을 담보할 수 있도록 대규모 인원을 동원하여 문장을 발화하거나 작성하도록 해 이를 DB화 하였

1) 중국어 음성인식 엔진의 경우 타 음성인식 엔진에 비해 동시 접속 사용자 수가 적은 것은 아직 서비스를 개시한지 오래 되지 않아 사용자 수가 비교적 적기 때문이다. 음성인식 성능을 높일 수 있도록 6가지 다른 LM에서 생성된 음성인식 이미지를 동시에 탐색하여 다중 음성인식 결과를 출력하고 있기 때문이기도 하다.

다. 이렇게 수집된 DB는 한국어의 경우 약 50만 문장에 달한다. 이는 여행/일상 상황을 가정하고 외국인간에 통역사를 통해 대화하거나, 같은 언어 사용자끼리 외국인과 대화하는 것으로 역할을 설정해 대화하거나, 또는 혼자서 외국인과 대화하는 시나리오를 작성하도록 해 수집한 것이다. 이 중 일부는 번역을 통해 다른 언어 음성인식 엔진의 언어 모델링에도 사용하였다. 두 번째 방법으로는 BTEC(Basic Travel Expression Corpus)[6]을 비롯해 여행, 일상 생활, 비즈니스, 항공, 호텔, 교통, 의료, 미용, 레스토랑, 스포츠 등 다양한 분야의 회화 예문에 대해 한국어의 경우 약 180만 문장을 수집하였다. 마지막으로는 드라마/영화 자막, 블로그, 일반 도서 등에서 약 540만 한국어 대화체 문장을 수집하였다. 지니톡의 경우 4년 이상의 기간 동안 연인원 6,000명 이상을 동원하여 다양한 문장을 수집하였다는 점과 여행/일상 대화 텍스트가 언어 모델의 중심을 이룬다는 점에서 다른 자동통역 시스템과 큰 차별성을 지닌다. 영어, 일본어, 중국어의 경우도 텍스트의 구성을 제외하고는 동일한 방법으로 구축한 후 이렇게 수집된 말뭉치를 이용하여 back-off 기반의 tri-gram 언어 모델을 구성하였다.

통상 음성인식엔진은 테스트 환경 및 평가 문장의 구성에 따라 성능에서 많은 편차를 보인다. 지니톡의 경우 현재 서비스 중인 음성인식 엔진은 사무실, 대로변, 골목, 식당 등에서 각 2,000발화씩의 대화체 여행/일상 영역 테스트 문장에 대해 평가를 실시한 결과 평균적으로 한국어 및 영어 음성인식 엔진은 90% 초반대의 단어 인식률을 보이고 일본어는 80% 후반대의 단어 인식률을 기록하였다. 중국어의 경우 사무실 환경에서는 80% 후반대의 단어 인식률을 보이나 아직 잡음 환경에서는 조금 취약한 상태이다.

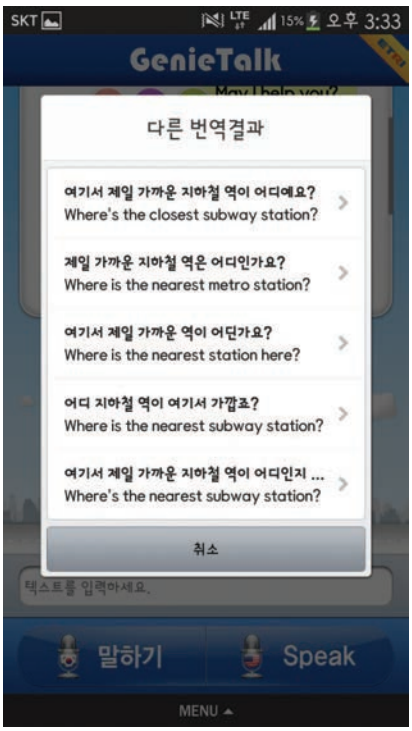
다. 자동번역 엔진

자동번역엔진의 경우 한/일, 일/한 자동번역의 경우 통계기반 자동번역 방법론을 도입하였고, 한/영, 영/한 및 한/중, 중/한 자동번역은 규칙 기반 자동번역 방법론을 채택하였다. 통계 기반 자동번역 방법은 언어 간의 유사성이 높을 경우 비교적 높은 번역 성능을 담보할 수 있고 또한 개발 기간을 단축할 수 있다고 알려져 있다. 이에 언어 간 유사도가 높은 한/일, 일/한 자동번역 엔진의 경우에는 대량으로 확보한 여행/일상 영역 대역 말뭉치를 중심으로 통계기반 자동번역 시스템을 구성하였다.[7] 이 때 Translation Model 용으로는 250만 대역 문장을 사용하였고 특히 이 중 189만 문장을 여행/일상 영역의 대화체로 구성하여 지니톡이 포커스를 맞추고 있는 여행/일상 분야의 대화체 번역에서 높은 성능을 나타낼 수 있도록 하였다. 또한 Language Model 용으로는 한국어 1,579만 문장 및 일본어 1,333만 문장을 사용하여 훈련하였고 이후 다양한 후처리 규칙을 도입하여 SMT의 문제점으로 지적되는 의문문, 부정문 등의 번역 오류를 해소할 수 있도록 노력하였다. 한 편으로 언어간의 유사도가 낮은 경우에는 신뢰도 있는 통계 정보를 획득할만한 말뭉치를 구하기가 쉽지 않고, 언어간의 특성이 달라 자동번역 과정에서 발생하는 오류 해소가 어렵다는 점에서 한/영, 영/한, 한/중, 중/한, 번역엔진의 경우 패턴 기반 자동번역 방법론을 채택하였다. 패턴 기반 자동번역의 경우 기존 규칙 기반의 자동번역기에 대용량의 문형 패턴 및 용언구 번역 패턴, 명사구 번역 패턴 등 다양한 패턴 지식을 추가하여 대화체에 특화하여 개발한 것이다.

이렇게 개발된 자동번역엔진은 각 언어쌍별로 여행/일상 분야에서 80~88% 정도의 번역 성능을 보인다. 음성인식엔진의 경우 결과가 틀리더라도 멀티모달을 이용한 수정 기능을 제공하므로 자동통역 성능은 자동번역엔진의 성능과 동일하다고 할 수 있다. 그러나 지니톡의 경우 다음 절에서 설명할 '다른 번역 결과'를 제공한다는 점에서 다른 자동통역시스템과 차별성을 가지고 있다. 번역 오류가 존재하는 문장에 대해 '다른 번역 결과' 검색을 실시할 경우 검색 성공률이 20.1%에 달하므로 실제 사용자들의 체감 성능은 더욱 높아지게 된다.

라. 지니톡의 부가 기능

지니톡은 음성인식 및 자동번역을 통한 자동통역 기



〈그림 5〉 다른 번역 결과 검색 화면

능 외에도 사용자의 편의를 고려한 다른 부가 기능들을 제공한다. 가장 주목할만한 것이 '다른 번역 결과'이다. '다른 번역 결과'는 음성 인식 문장에 대해 자체적으로 수집한 2백만 문장 이상의 대용량 예문 데이터베이스를 검색해 그 결과를 표시하는 기능이다. 인식된 문장에 '다른 번역 결과'가 존재할 경우 검색된 예문의 숫자가 인식된 문장 옆에 아이콘으로 표시되고 아이콘을 터치하게 되면 〈그림 5〉처럼 '다른 번역 결과' 검색 예문이 표시된다.

'다른 번역 결과'는 단순히 TM(Translation Memory)을 검색한 결과가 아니고 입력문에서 주요 키워드를 추출한 다음 키워드 별로 가중치를 부여하고 이를 이용해 유사도별 랭킹 알고리즘을 통해 기번역된 예문DB를 검색한 결과이다. 또한 고유명사의 경우 '인명', '상호명', '지명' 등의 해당 속성 기반으로 클래스화하여 대표 검색을 실시함으로써 검색 성공률을 높이도록 하였다.

그리고 또 다른 기능으로 통역 대상 외국어가 익숙치 않은 사람들을 위해 지니톡 메인 화면에서 '합성음 듣기 & 발음 보기' 기능을 제공한다. 이 기능을 통해 통역된 문장을 다시 듣고 따라 읽어볼 수 있어 합성음만이 아니라 자신의 목소리로 상대방에게 본인의 의사를 전달할 수 있게 된다. 특히 이 기능은 해당 발음을 한글을 이용해 표시하고 있어 중국어, 일본어와 같이 한자 또는 가나로 표기되어 문자 자체를 읽을 수 없는 경우매우 큰 도움이 된다.

이 외에도 지니톡의 UI 곳곳에는 사용자의 경험을 고려한 디자인이 내재되어 있다. 먼저 메인 통역 화면도 통상의 SMS, 메신저와 유사한 형태로 설계해 사용자가 익숙하게 히스토리 기반의 대화형으로 자동통역기를 이용할 수 있도록 하였다. 또 지니톡을 처음 사용하는 이용자들을 위하여 친근한 캐릭터를 등장시켜 사용법에 대해 말풍선으로 가이드함으로써 지니톡의 다양한 기능들에 쉽게 접근할 수 있도록 하였다. 그 외에도 사용자가 자주 쓰는 예문들을 미리 저장해 놓고 편리하게 이용할 수 있는 북마크 기능, 음성인식에 실패했을 경우에도 손쉽게 음성인식 결과를 수정할 수 있는 문장 편집 기능, 자동통역 결과를 타 앱에서 활용할 수 있는 문장 복사 기능 등도 제공하고 있다. 특히 사용자가 발견한 자동통역 오류를 바로 연구자에게 피드백할 수 있는 오류 신고 기능을 채택해 사용자의 피드백을 자동통역 시스템의 성능 개선에 활용하고 있다. 현재까지 기록된 사용자의 피드백은 24만 여건에 이른다.

대화형 영어 학습 서비스인 지니튜터는 원어민 교사 없이 컴퓨터를 이용하여 영어 말하기 연습을 수행하며, 영어 표현 학습, 문법 및 발음 교정 등을 제공받을 수 있는 음성언어 기술 기반 영어 학습 서비스 이다.

지금까지 지니톡을 통한 대국민 자동통역 시범 서비스에 대해 알아보았다. ETRI에서는 이렇게 시범 서비스를 통해 확보한 자동통역 로그를 기반으로 자동통역 원천 기술을 꾸준히 연구해 나가는 한편 자동통역 대상을 스페인어, 프랑스어 등 유럽어권까지 확대해 국민편익 증진 및 국민 화합을 통한 국가 경쟁력 확보와 기

업의 수출 경쟁력 강화에도 기여해 나갈 예정이다.

2. 음성인식 영어학습 서비스

대화형 영어 학습 서비스인 지니튜터는 원어민 교사 없이 컴퓨터를 이용하여 영어 말하기 연습을 수행하며, 영어 표현 학습, 문법 및 발음 교정 등을 제공받을 수 있는 음성언어기술 기반 영어 학습 서비스 이다. 이 서비스는 한국전자통신연구소에서 개발되었으며 현재 시제품 단계에 있다. 본 서비스는 영어 학습 요소 중 말하기 학습에 중점을 두고 개발하였으며, 한국인 영어 발음 특성을 고려한 발음 클리닉과정과 다양한 상황에서의 영어 말하기 표현 및 대화 문장을 학습하는 Think & Talk 과정, 주어진 사진이나 그래프를 보고 말하기 표현을 학습하는 Look & Talk 과정으로 이루어져 있다.

가. 한국인 영어 발음 특성을 고려한 발음클리닉

한국어 발음 특성과 한국인의 영어 발음 특성을 고려하여 발음클리닉이 총 30개의 레슨으로 구성되었으며, 각각의 소리를 학습하는 basic level, 한국인이 유의해서 발성해야할 발음을 학습하는 intermediate level, 한국어와 다른 조음현상을 배우는 advanced level 로 나뉜다.

발음클리닉은, (a) 학습할 소리 직접 발성 (b) 동영상 강의 (c) 각 소리/단어/문장에 대한 음성인식 기반 학습 (d) 학습한 음성 재생 (e)학습한 소리에 대한 문장 단위 평가로 진행 된다.

음성인식 기반 학습이 진행되는 발음클리닉을 위하여, 한국인 영어 음향모델을 이용한 음성인식 적용, 각 음소 및 학습 음소에 대한 발음 평가 점수 획득, 발성 단어 및 문장에 대한 억양 점수 획득, 원어민 발성 음성과의 비교를 통한 학습자 음성에 대한 점수 계산 등의 기술들이 적용되었다.

나. 영어 표현 및 대화학습

영어 표현 및 대화 학습 과정인 Think & Talk, Look & Talk 교육 과정은 영어 연속어 음성인식 기술, 영어 대화 처리 기술 및 언어 처리 기술을 적용하여 개발되었다. Think & Talk 은 대화의 난이도, 문장의 난이도를 고려하여 Level 1, 2로 나누어지며, 각 6개의 레슨으로 구성되어 진다. 각 레슨에서는 특정한 주제, 상황이 주어지고, 해당 주제 상황에의 다양한 조건을 학습자가 선택하여 학습자가 다양한 문장 표현을 학습할 수 있도록 지원하고 있다. Look & Talk 은 난이도를 고려하여 총 11개의 레슨으로 구성되어 진다. Look & Talk 과정의 레슨에서는 사진, 그래프, 팜플렛 등이 주어지고 해당 이미지에 대해 설명할 수 있는 문장 표현을 학습하도록 구성되어져 있다.

각 레슨은 (a) 핵심 표현 학습(Key express) (b) 오디오 강의(Lecture) (c) 프리뷰(Preview) (d) 문장 연습(Practice) (e) 대화 학습(Challenge) 단계 로 진행된다. 프리뷰 단계에서는 대화를 진행하기 위한 조건 값을 학습자가 선택하게 되며, 학습자는 학습할 때마다 서로 다른 조건을 선택함으로써 다양한 문장 표현을 학습할 수 있게 된다. 문장 연습 단계에서는 주어진 대화를 구성하는 가장 적합한 문장을 원어민 발성을 듣고 따라서 말하기 연습을 하는 Listen and Repeat와 Role Play를 연습을 할 수 있다. 마지막 대화 학습 단계에서는 시스템 질문에 학습자가 적절히 대답하고, 이에 대한 지니튜터의 피드백을 받으며 대화를 진행해 나가며 학습을 진행한다.

Think & Talk 학습 과정은 특정 주제에 대해 시스템이 질문을 던지고, 학습자는 이에 대한 대답을 함으로써 영어 표현 능력을 학습한다. 학습자가 발성한 영어 문장을 연속어 음성인식 기술을 적용하여 인식을 수행하고, 인식된 문장에 기반하여 학습자 문장의 내용이 적합한지, 문법적 오류가 있는지, 표현적인 오류가 있는지를 분석하고 이에 따라 적절한 피드백을 제공한다. 학습자가 문법적으로 잘못한 부분에 대한 피드백을 제공하고, 적절한 정답 표현들을 제시하여 주어진 상황에서 적절한 문장 표현을 학습할 수 있도록 교육한다.

Look & Talk 학습 과정은 인물, 장소, 매뉴얼, 바 그래프, 파이 그래프가 주어지고 학습자는 해당 이미지에 대한 시스템의 질문에 대해서 대답함으로써 적절한

영어 표현 방법을 학습하게 된다.

다. 한국인 영어에 강인한 영어 연속어 음성인식 기술

지니튜터는 영어 대화 학습을 진행하기 위하여 영어 연속어 음성인식을 수행한다. 그러나, 영어 학습자의 경우 부정확한 영어 발음과 불완전한 문법으로 발성을 하게 된다. 이에 한국인 영어에 강인하게 인식할 수 있는 음성인식 기술이 적용되어야 한다.

한국인 영어에 강인한 음성인식을 수행하기 위하여 한국인이 발성한 영어 문장 DB를 구축하여 음향모델 학습에 반영하였다. 또한 한국인이 쉽게 유발하는 발음 오류를 모델링하여 한국인 영어에 맞는 발음사전을 생성하였다.

영어 학습자가 유발하는 문법적 오류에 강인한 음성인식을 수행하기 위하여 학습자 오류를 반영한 언어모델을 학습하였다. 문법 오류는 명사의 단복수 오류, 동사의 인칭 오류를 비롯하여 다양한 오류를 포함한다. 주어진 도메인에 대해 자주 유발하는 학습자 오류는 오류 메모리로 모델링 하고 이에 대한 오류 코퍼스 역시 생성하여 학습에 반영한다. 이와 같은 과정을 통해 학습자가 흔히 유발하는 문법적 오류를 포함하는 언어모델을 학습하였다.

라. 시각정보 기반 영어 회화 유도 대화 기술

비원어민 학생의 경우 영어 대화상황에서 어떠한 내용을 말해야 할지, 그리고 어떤 표현을 써야 할지 몰라서 회화에 어려움을 겪는 경우가 많다. 본 연구에서는 대화 영어 회화를 유도하기위해 대화의 중심이 되는 주제들을 사전에 모국어로 검토 및 결정하도록 하여 내용적인 측면에서 교육 가이드를 주는 방법과 대화 중에 나타나는 표현 오류를 바탕으로 어떤 표현을 해야 더 적절한가를 계속해서 피드백해 줌으로써 대화 끝까지 발화를 이어나갈 수 있도록 유도하는 방법을 구현하였다.

〈그림 6〉에서 보듯이, 특정 대화 주제를 선택하게 되면 구체적으로 어떠한 내용을 가지고 대화를 진행할 것인지를 모국어로 선택하거나 그림으로 표현되어 어떠한 내용으로 대화할지 직관적으로 알 수 있게 하고 대화

〈그림 6〉 시각정보 기반 대화 유도 예시

〈그림 7〉 지니튜터 학습시스템 피드백 예시

중간에 말할 내용이 표시되어 대화의 흐름을 잊지 않도록 유도한다.

대화 중간에서 구체적인 표현을 모르거나 오류가 있는 부적절한 발화를 하였을 때 시스템이 상황에 더 잘 맞는 표현을 가이드 해줌으로써 학습자가 적절한 표현을 사용하여 대화를 쉽게 진행할 수 있도록 가이드 한다. 〈그림 7〉을 보면, 학습자가 단순히 "a rocking

char" 라고 답변하였을 때, 영어교육 대화 시스템이 구체적이며 완결된 형태의 문장 표현을 제시해 줌으로써 대화를 진행할 수 있도록 가이드 한다.

마. 대화 시스템 평가

영어 교육 대화 시스템을 평가하기 위해 다음과 같은 실험자, 태스크, 측정 기준을 사용하였다.

- 실험자: 30명으로 영어 토익 실력을 토대로 초급:중급=1:1의 비율로 선정하였다(초급은 토익 500점 이하, 중급은 토익 501~900점 이하로 선정하였다)
- 태스크: 실험자는 Entertainment, Food, Work, People, Place의 5개 도메인에 대해 지니 튜터와 대화를 실시한다.
- 측정 기준: 평가자는 실험자의 발화 로그를 대상으로 다음과 같은 측정 기준에 의거 평가를 실시하였다.
  - 태스크 성공률(Task Success Rate)=성공한 태스크수/전체 태스크수
  - 대화턴 성공률(Turn Success Rate)=정확한 시스템 반응수/전체 시스템 반응수
  - 오류 정정률(Error Correction Rate)=시스템의 정확한 오류 수정수/오류가 있는 사용자 발화수

상기와 같은 측정 기준을 토대로 영어 교육 대화 시스템을 평가한 결과, 태스크 성공률은 90.67%였으며 대화턴 성공률은 92.16%였다.

3. 대화형 내비게이션 시스템

차량 환경은 음성 인터페이스가 효과적으로 사용될 수 있는 대표적인 환경 중의 하나로, 오래 전부터 많은 연구가 이뤄져왔고, 따라서 많은 종류의 음성 인터페이스 제품이 이미 상용화되어 널리 쓰이고 있다. 지금까지의 차량 환경 음성 인터페이스 제품은 대부분 단말 내장형으로 비교적 단순한 형태의 명령어 인식 또는 목적지 인식이 주를 이루어 왔다.

최근 들어 여러 어플리케이션에서 음성 인식 성능이 향상됨에 따라 음성 인터페이스에 대한 사용자 기대치가 높아졌으며 이에 따라 보다 종래의 명령어 위주가 아니라 사람과 대화하는 것과 유사한 방식의 음성 대화를 통하여 사람-기계간 인터페이스를 만들고자하는 수요가 급증하였다. 또한 스마트폰의 빠른 보급과 함께 차량 환경에서도 인터넷 연결이 가능한 경우가 많아짐에 따라 네트워크를 통한 서버-클라이언트 방식의 음성인식이 차량에서도 가능하게 되었다. 이러한 환경에 발맞추어 본 과제에서는 차량환경에 특화된 서버-클라이언트 기반 차량환경 대화형 내비게이션 서비스 시스템을 개발하였다.

대화형 내비게이션 서비스 시스템은 사용자가 원하는 정보를 사용자가 사람에게 물어보는 것처럼 대화를 통하여 질문을 하고 시스템이 대화를 이해하여 사용자가 원하는 정보를 제공하는 것을 목적으로 한다.

대화형 내비게이션 서비스 시스템은 사용자가 원하는 정보를 사용자가 사람에게 물어보는 것처럼 대화를 통하여 질문을 하고 시스템이 대화를 이해하여 사용자가 원하는 정보를 제공하는 것을 목적으로 한다. 예를 들어 사용자가 '서울에 있는 국립중앙박물관을 찾아줘' 라고 발성하는 경우 '서울' 지역에 있는 '국립중앙박물관' 이라는 지명을 찾는다는 사용자의 의도를 파악하고, 해당하는 지명을 목적지로 제시하게 된다. 서비스를 대상으로 하는 영역은 〈표 1〉과 같은 다섯 개의 영역으로 제한하지만, 음성인식의 경우 이외의 영역에 대해서도 딕테이션 수준의 음성인식 성능을 제공할 수 있도록 구성한다.

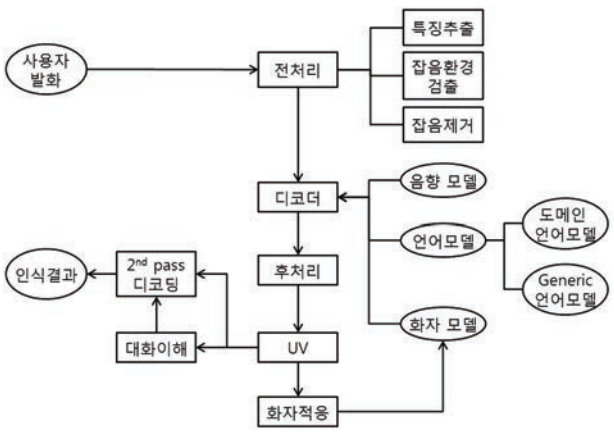
〈표 1〉 대화형 내비게이션 시스템의 서비스 영역

영역	예
경로설정, 주변검색	이마트를 찾아줘, 우리집으로 가자, 가장 가까운 주유소 안내해, 하나은행 경유해 등
경로 정보	얼마 남았지, 목적지까지 남은 시간은 등
교통정보	올림픽대로 막히나, 현재 경로 앞에 사고가 있나 등
날씨정보	대구 날씨는, 내일 서울에 비가 오나, 오늘 운동해도 좋을까 등
DMB	MBC 틀어줘, 프로야구 보여줘, 등
명령어	음악 틀어줘, 음악 꺼, 지도 띄워줘 등

가. 대화형 내비게이션 서비스용 음성인식 기술

대화형 내비게이션 서비스 시스템을 구성하기 위해서는 우선 차량 환경에 특화된 음성 인식 시스템이 필요하다. 대화 인터페이스의 전체적인 성능은 내장된 음성 인식 시스템의 성능에 좌우되므로 높은 성능의 음성 인식기를 구현하는 것이 중요하다. 차량 환경에 특화된 음성 인식 시스템의 기능별 구조도를 〈그림 8〉에 나타내었다.

전처리 모듈은 음성 신호를 입력으로 받아서 인식에 필요한 특징 벡터를 추출하는 모듈로 다음과 같은 기능을 수행한다. 첫 번째 기능은 잡음 제거 기능으로 입력 신호에서 잡음을 제거한다. 일반적으로 차량의 주요 잡음원인 주행 및 공조 잡음은 정적 잡음으로 Wiener 필터가 잘 동작하여 잡음 환경에서의 음성 인식 성능은 크게 개선되나, 정차 환경에서의 음성 인식 성능이 저하되는 문제점이 있다. 이는 조용한 환경의 경우 Wiener 필터가 잡음제거보다는 음성 왜곡에 더 큰 영향을 미쳤음을 의미한다. 이러한 문제점을 해결하기 위하여 신호의 앞부분을 이용하여 현재 발성 환경이 잡음 환경인지 아닌지를 검출하는 자동 잡음 제거 기능을 추가하였다. 개선된 잡음 제거 알고리즘에서는 음성 신호의 초기 일정 신호를 분석하여, 신호 레벨이 일정 문턱값보다 크면 잡음 환경으로 간주하여 Wiener 필터를 적용하고, 그렇지 않으면 Wiener 필터를 적용하지 않는다. 이와 같은 방법으로 주행시 인식 성능과 조용한 환경에서의 인식 성능을 동시에 높일 수 있다. 두 번째로 입력된 신호 중 발성구간을 검출하는 끝점 추출기능으로 종래의 알고리즘을 적용하였다. 세 번째 기능은 입력 신호를 인식에 적합한 특징 벡터로 변환하는 특징 추출 기능이다. 음성인식에 위한 특징으로는 새롭게 개발된 53차 MFCC를 적용하였다[8].



〈그림 8〉 대화 음성 인식 시스템 구조도

대화형 내비게이션 서비스 시스템을 위한 음성 인식용 음향 모델은, POI 인식과 함께 정보 조회 목적의 대화형 음성 인터페이스를 위해 연속어 음성 특성이 반영되도록 훈련용 음성 DB를 구성하였다. 훈련용으로 사용된 음성 DB의 구성은 딕테이션 영역 음성 DB 750 시간 분량과 정보검색 로그, 숫자음 및 POW 350 시간 분량으로 총 약 1,100시간 분량의 음성 데이터 및 대응되는 전사문으로 구성되었다. 1차 음향 모델은 SMS/트위터 등 모바일 환경에서의 딕테이션 인식에 사용한 일반적인 음향 모델로 주로 조용한 환경에서의 음성 인식을 대상으로 구축되었다. 잡음 환경에서의 음성 인식 성능을 개선하기 위하여 1차 음향모델을 기본으로 자동차 환경 잡음 신호를 가산하여 음향 모델을 훈련하여 2차 음향 모델을 구축하였으며, 이후 다중 공간 GMM 기반의 음향 모델을 구성하여 정차 환경과 주행환경에서의 음성 인식 성능을 동시에 최적화하는 2 차 음향 모델을 구축하였다. 여기에 추가로 주행 환경에서의 성능 개선을 위해 실제 자동차 주행 환경에서 수집된 8 시간 분량의 소량의 음성 DB를 이용하여 최대 사후 추정확률 방식으로 적응하여 최종적인 음향 모델을 구축하였다. 적응 훈련은 소량의 음성 DB로도 실제로 서비스하는 환경에서 큰 성능 개선을 가져올 수 있다는 장점이 있다.

대화형 내비게이션 시스템 개발에서 가장 어려운 점 중의 하나는 기존의 자유 발화 인식 기능에 덧붙여 내비게이션 사용을 위한 목적지명이 인식이 되어야 한다는 점이다. 목적지명은 그 개수가 300만개 이상으로 매우 많고 또한 일반적인 대화에서는 사용되지 않는 단어들로 이루어진 고유 명사의 빈도가 높다. 이러한 단어들이 포함된 코퍼스가 절대적으로 부족한 상황에서 언어 모델 생성을 위한 단어의 출현 확률을 추정하는

것이 매우 어려운 작업이다. 이러한 어려움을 해결하기 위하여, 대화형 내비게이션 시스템을 위한 음성 인식 시스템은 두 가지 종류의 언어모델을 적용하였다. 첫번째는 'OOOO 검색해줘' 와 같이 일반 사용자가 내비게이션 사용 환경에서 높은 빈도로 사용할 것 같은 문형 내에서 인식하는 언어 모델로 JSGF 문법을 이용하여 구축되었다. JSGF 문법을 이용할 경우 그 문형을 사전에 쉽게 설계하고 추가할 수 있다는 장점이 있다. 또한 사용자가 해당 문법에 정확히 일치하도록 발성한 경우 더 높은 인식 성능을 갖는다. 둘째는 이러한 문형을 벗어나서 비교적 높은 자유도로 발성하는 경우를 인식하기 위한 언어모델로 300만개 이상의 목적지명을 12만 형태소로 나누고 정보조회 서비스 및 일반 딕테이션 영역의 어휘를 포함하여 서비스 영역에 맞도록 최적화된 n-gram 언어 모델을 생성하였다. N-gram 방식의 언어 모델은 어떠한 형식의 입력이든 비교적 높은 성능의 인식 성능을 보이는 반면, JSGF 방식의 언어 모델은 정해진 입력에 대해서는 높은 성능을, 그렇지 않은 입력에 대해서는 낮은 성능을 보인다. 대화형 내비게이션 서비스를 위해서는 두 가지 특징이 필수적이므로, 실제 시스템에서는 두 언어모델 기반의 인식기를 동시에 수행하여, 신뢰도를 계산한 후 신뢰도가 높은 결과를 선택하도록 하였으며, 두 인식기의 특성을 반영하여 적절한 수준의 가중치를 두어 최적의 인식 성능이 나오도록 설정하였다.

스마트TV를 위한 음성인식 기술은 삼성전자 및 LG전자 등 주요 가전사뿐만 아니라 구글 및 다음커뮤니케이션 등과 같은 인터넷 포털, 애플 및 마이크로소트와 같은 컴퓨터/소프트웨어 업체 등에서도 관심을 갖고 전략적 사업화를 도모하고 있는 주요한 핵심 기술 중의 하나이다.

대화형 내비게이션 시스템은 차량 환경에서 운전자가 사용하는 경우가 대부분으로, 일반적인 경우 1명의 사용자만 사용하므로 화자 적응 기법을 적용하기에 적합한 응용 분야이다. fMLLR 방식의 화자 적응 알고리즘이 사용되었으며 각 화자별로 인식이 정상적으로 수행된 경우에 한하여 인식 결과 및 입력 특징벡터를 이용하여 발화 단위로 화자 적응을 적용하였다. 또한 앞서 설명한 바와 같이 대화형 내비게이션 시스템을 위한 음성 인식에서 어려운 점 중의 하나는 수 백 만개의 목적지명을 인식해야하는 것이다. 이러한 문제를 해결하기 위하여 1차 음성 인식 결과로부터 대화 이해 모듈의 결과를 도출하고, 대화 이해 모듈의 분석 결과 사용자가 찾고자 하는 정보가 목적지명에 해당하는 경우 이 구간의 음성을 재인식하는 과정을 거치게 된다. N-best인식할 대상이 정해지면 먼저 저장된 특징 벡터열로부터 모노폰 열을 인식하고, 인식된 모노폰 열로부터 lexical 탐색을 통하여 인식 대상 어휘를 선정한다. 이후 선정된 대상 어휘로부터 트라이폰 인식을 수행하여 rescoring하여 최종적으로 N개의 인식결과를 도출해낸다.

나. 대화형 내비게이션 서비스를 위한 대화처리 기술

앞서 설명한 5개 영역의 차량용 내비게이션 정보서비스에 필요한 대화모델을 위해, 47개의 대화의도 타입과 의미망을 정의하였다. 〈표 2〉는 각 분야의 대표적인 대화의도들을 보이고 있다.

정의한 도메인에서 사용자의 대화를 이해하고 정보를 서비스하기 위한 의미망은 먼저 크게 클래스(class)와 각 클래스의 자질을 표현하는 슬롯(slot)으로 구성된다. 클래스는 상하위 관계를 가질 수 있으며, 각 자질은 문자열, 수 등이나 클래스를 값으로 가질 수 있다. 본 과제에서는 클래스는 대문자로 시작하는 영어 단어로 기술하고, 슬롯은 소문자로 구성된 영어 단어로 기술하며

〈표 2〉 분야별 대화의도 예

관련 분야	대화의도 타입
경로 설정, 주변 검색	introduce
경로 정보	request(time), request(distance) ...
교통 정보	request(Traffic.info)
날씨 정보	request(Weather.info) ...
DMB	play
명령어	execute, end

소속된 클래스명을 앞에 붙여 다른 클래스의 유사 슬롯과 구분한다. 차량용 정보서비스용 한국어 대화모델에는 6개 도메인에 대해 상기 10개 클래스와 100 슬롯으로 구현하였다. 대화모델에 필요한 사용자 발화를 14,000여 발화를 수집하고 상기 대화의도와 의미망으로 태깅하였다. 차량용 정보서비스용 한국어 대화모델에서 대화관리를 위한 계층 태스크 그래프의 경우 복합 태스크로 구성되지만 각각 태스크의 복합적 작용이 적어서 태스크 간의 순차적 나열이 적도록 최적화하게 구성하였다. 또한, 계층 태스크 그래프의 태스크에 수행할 시스템 대화에 대해서는 대화라이브러리를 구축하였다.

4. 스마트 TV 음성 인터페이스

최근 수년동안 스마트 기기 시장의 폭발적인 증가에 따라 스마트 기기에 채용되는 음성인식 기술도 점차 그 사용영역이 확대되고 이에 따른 시장도 확장 일로에 있다. 스마트TV를 위한 음성인식 기술은 삼성전자 및 LG전자 등 주요 가전사 뿐만 아니라 구글 및 다음커뮤니케이션 등과 같은 인터넷 포털, 애플 및 마이크로소트와 같은 컴퓨터/소프트웨어 업체 등에서도 관심을 갖고 전략적 사업화를 도모하고 있는 주요한 핵심 기술 중의 하나이다.

〈그림 9〉는 한국전자통신연구원에서 제시한 차세대 스마트TV 개요도로서, TV와 인터넷의 결합을 기반으로 이용자 친화적인 멀티모달 인터페이스에 의해 제어가 가능하고, N-스크린을 기반으로 방송, 통신, 컴퓨터형 서비스를 제공하는 CPND(Contents, Platform, Network, Device) 기반의 생태계로 요약된다.

〈그림 9〉 차세대 스마트TV 개요도

〈표 3〉 음성 인식과 동작 인식의 장단점 비교

구분	음성인식	동작인식
장점	• 복잡한 명령 가능 • 정보검색 등 지식서비스에 적합 • 어두운 곳에서 사용 가능	• 원거리 제어가 가능함 • 소음이 심한 환경에서도 가능함
단점	• 소음이 심한 환경에서 사용 어려움 • 원거리에서 사용 어려움	• 단순한 명령만 가능 • 어두운 곳에서 사용 어려움

삼성전자와 LG전자는 음성인식 전문업체인 Nuance의 서버형 솔루션을 통해 음성 및 동작 인식을 포함하는 멀티모달 인터페이스를 자사의 기본 UI/UX로 기본 적용하고 있다. 구글은 자사의 크롬, 유튜브, 구글 플레이 등을 통합하여 서비스하는 구글TV 플랫폼을 공개하고 독자적으로 구글 셋탑을 출시하는 한편 가전사와의 제휴 모델도 공개하고 있다. 마이크로소프트는 음성인식과 Kinect 등 제스처 인식 모듈을 통합한 셋탑인 Xbox를 공개하여 멀티모달을 강조한 UI/UX를 강화하는 사업화 전략을 취하고 있다. 애플은 스마트폰상에서 이미 서비스 중인 대화형 개인비서인 Siri를 기반으로, 2014년 출시 예정인 애플 iTV 플랫폼에 기본 UI/UX로서 음성인식 인터페이스를 제공할 예정이다.

주요 멀티모달로 사용되는 음성 및 동작 인식의 주요한 장단점을 살펴보면 〈표 3〉과 같으며, 이들 두 모달리티는 상호보완적으로 적용할 경우 사용자 편의성을 극대화할 수 있다.

가. 스마트TV용 음성인식 기술

스마트TV에 적용되는 음성인식 기술은 통상 서버/클라이언트 방식과 내장형 방식으로 구분할 수 있다. 서버의 경우 충분한 계산 자원을 기반으로 복잡한 자연어형 질의-응답이나 대화형 서비스가 가능하며, 내장형의 경우 제한된 계산 자원에 따라 명령/제어를 위한 소규모 어휘나 간단한 검색 서비스를 지원 가능하다. 이러한 서비스를 요약하면 〈표 4〉와 같다.

스마트TV를 위한 음성인식 적용 시 고려 사항은 가장 먼저 TV 스피커에서 출력되는 음향효과, 가정 환경

에서 일상적으로 발생하는 생활 잡음을 가정한 잡음 처리 기술이다. 즉 TV에서 출력되는 소리와 생활 잡음 등이 발성자의 목소리와 혼합되어 마이크로폰에 입력됨으로써 음성인식 오동작을 일으키게 되는 것이다. 이를 위해 적응형 잡음 제거(adaptive noise cancel) 기술이 필수적으로 요구된다.

마이크로폰을 사용해야만 하는 음성인식의 기술 특성상 입력 장치로서 스마트폰이나 마이크로폰이 탑재된 리모트 콘트롤 장치를 사용하는 것 외에 TV 단말에 마이크로폰을 탑재하는 방식이 있다. 이때 보통 TV와 사용자와의 거리는 3~5미터 또는 그 이상 격리되어 있으므로 사용자 음성이 마이크로폰에 도착되면서 급속이 음압이 감쇄되면서 주변 소음의 영향을 많이 받을 수밖에 없는데 이를 위해 보통 마이크로폰 어레이를 사용한 빔포밍 기법을 이용하거나 하여 고도의 지향성 입력 조건을 구성해야 한다.

하드웨어 기술의 비약적 발전에 따라 음성인식 엔진에서 적용하는 어휘는 아직 완전하지는 않지만 음성검색이나 받아쓰기가 가능할 정도의 기술 혁신이 이루어지고 있다. 하지만 등록되지 않은 어휘는 인식이 불가한 음성인식 기술의 특성과 수시로 변화하는 프로그램 정보의 특성을 고려하면 방송사의 방송 정보 또는 VOD 콘텐츠 정보를 실시간으로 반영하여 음성인식용 리소스를 구성해야만 한다. 이를 위해 음성인식 엔진은 대용량 어휘에 따라 인식용 문법 네트워을 실시간 또는 최단시간내에 갱신할 수 있도록 하는 메카니즘이 필수적이다.

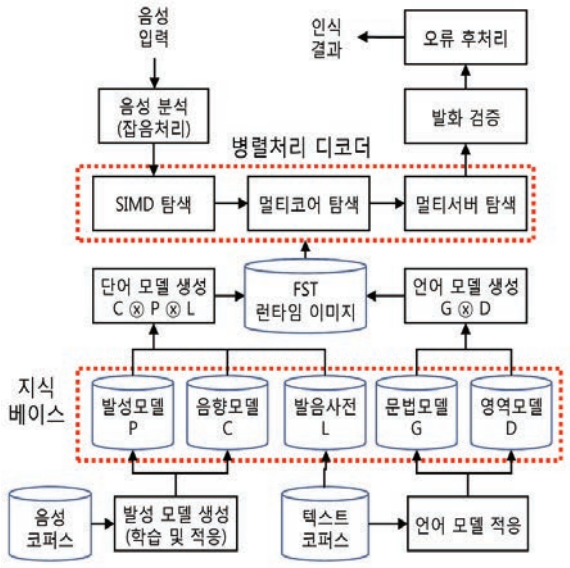
방송 정보 서비스를 위해 프로그램명에 대한 이형태 생성이 필수적인데 예를 들어 "해를 품은 달"이라는 방송물에 대해서 일반인들의 검색 패턴은 "해품달" 등과 같이 첫음절 만을 적용하거나 심하게 축약된 검색이 일반화되고 있다. 이를 위해 방대한 양의 방송 콘텐츠를 대상으로 음성인식용 어휘를 자동으로 추출하는 등의 이형태 구축이 필요하다. 이를 위해 사전이나 패턴 기반의 생성 방법이 필요한데, 예를 들어 "프리즌 브레이크 시즌 1의 17회"등에서와 같이 보통 회차 정보, 제목, 소제목 등의 프로그램명 구성 단위를 인식하고 활용하여 "프리즌 브레이크", "프리즌 브레이크 17회", "프리즌 브레이크 시즌 1" 등과 같은 이형태 후보들을 생성하여 사용하는 것이다.

〈표 4〉 서버형 및 내장형 음성인식기술의 적용성 비교

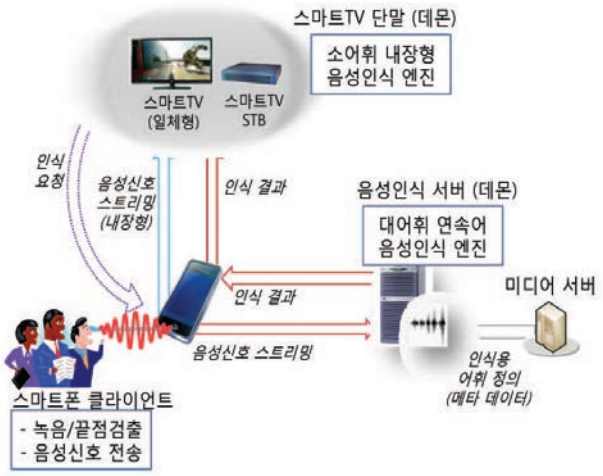
구분	특징	서비스
서버형	• 풍부한 계산 자원으로 제한 없는 서비스 가능 • 단말 수요에 따른 서버 비용 증가	• 자연어 질의형 문장 인식 • 대화형 서비스, 매쉬업 서비스 • 음향/언어 모델 학습에 따른 음성인식 성능의 지속적 향상 가능
내장형	• 저속CPU 및 자원제약으로 인한 서비스 제약 • 네트워을 사용치 않아 저가형 서비스 가능	• 명령/제어 기능 : TV제어, 앱 구동, OSD 제어 • EPG 검색 : 프로그램, 출연자, 장르 검색 가능

가정이나 기타 장소나 공히 스마트TV는 연령, 성별과 관계없이 다양한 사용자가 사용하게 된다. 이러한 환경에서 개별 사용자의 음성 인식률을 제고하기 위해서는 화자의 음성을 누적하여 사용할수록 성능이 개선되는 효과를 보기 위해 일반적으로 적응형 화자 음성 학습 기술이 적용된다.

음성 검색이나 받아쓰기와 같이 언어 모델 구성을 위한 충분한 데이터가 존재하는 영역과 달리 TV 환경에서의 음성 입력 패턴이나 양태에 대해서는 별도로 사용가능한 텍스트 코퍼스가 존재하지 않게 된다. 이를 위



〈그림 10〉 음성인식 서비스를 위한 기술 구성도



〈그림 11〉 음성인식 서비스를 위한 시스템 흐름도

해 사용자 양태 분석 및 전용 언어 모델 구성을 위한 코퍼스를 구축하고 이를 태깅하여 음성인식용 언어 모델을 구성해야 한다.

상기의 적용 기술을 고려한 음성인식 엔진의 런타임 리소스를 생성하고 이를 실시간으로 적용하는 과정을 도해하면 〈그림 10〉과 같다[9]. 음성인식에서 사용하는 지식은 크게 "단어 모델"과 "언어 모델"로 구분할 수 있는데 단어 모델은 비교적 고정적으로 사용되는 지식이며 언어 모델은 동적으로 변경되어 지는 지식이다. 다양한 화자의 다양한 음성 정보를 고려하여 오프라인으로 학습되어 저장된 음향 모델과 발성 모델은 콘텐츠 또는 EPG 서버로부터 확보되는 동적 프로그램 정보와 결합되어 일차적으로 단어 모델을 생성한다. 또한 이와 같은 수준에서 미리 정의되어 있는 문법/영역 모델에 따라 언어 모델을 생성하게 된다. 최종적으로 단어 모델과 언어 모델을 통합하여 FST 기반의 런타임 이미지를 생성한 후 음성인식 엔진의 런타임 콤포넌트인 디코더에서 이를 사용하여 인식을 수행하게 된다[10].

나. 스마트TV 음성인터페이스 시스템의 구성도

〈그림 11〉은 음성인식 엔진에 기반하여 실제 음성 인터페이스를 구성하는 서비스 시스템을 소개하고 있다[9].

〈표 4〉에 따라 내장형 엔진은 명령/제어나 고립어형태의 프로그램명 검색을 수행하고, 서버는 대어휘로 구

〈표 5〉 서비스 흐름

1. 미디어 서버로부터 EPG 정보를 수신하여 음성인식용 리소스를 빌드한다.
2. 스마트폰을 통해 녹음된 음성 신호를 서버로 실시간 음성 스트리밍을 수행한다.
3. 음성인식을 수행한다.(서버 및 내장형 별도)
4. 음성인식의 인식결과를 생성하여, 서버의 인식 결과를 클라이언트로 전송한다.
5. 음성 인식 결과에 따라 EPG 검색 및 기기 제어 등의 서비스를 수행한다.

성되는 자연어 형태의 음성을 입력받아 서비스를 수행하는 과정이다.〈그림 11〉에 따른 대략의 서비스 흐름을 정의하면 〈표 5〉와 같이 요약할 수 있다.

V. 결 론

지금까지 음성언어처리 분야의 전반적인 기술 동향과 사업화 현황에 대하여 살펴보았다. 또한 국내의 한국전자통신연구원에서 개발한 대표적인 서비스에 대하여 간단히 살펴보았다.

음성언어처리기술 관련 산업은 전 세계적으로 볼 때 아직은 초기 단계라 할 수 있으나 최근 들어 나타나기 시작한 다양한 서비스와 앞으로 전망되는 지식서비스의 성장 가능성을 고려할 때 본 기술의 중요성은 매우 높다고 할 수 있다.

특히, 본 기술은 언어와 직접적인 관련이 있는 기술이므로 독자적 언어체계를 가진 우리나라에서는 한국어에 대한 기술 경쟁력을 유지하고 발전시킴으로써 향후 한국어 기반의 지식서비스가 세계적 경쟁력을 가질 수 있는 기반을 만들어야 하겠다.

참 고 문 헌

[1] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Goole Search by Voice: A Case Study," in Visions of Speech: Exploring New Voice Apps in Mobile Environments, Call Centers, and Clinics, A. Neustein, Ed. Springer, 2010

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohaned, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.

Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, Vol. 29, No. 6, pp. 82-97, 2012

[3] 나승훈 외, "CRF에 기반한 한국어 형태소 분할 및 품사 태깅", 한글 및 한국어정보처리, 2012

[4] 나승훈 외, "래티스상의 구조적 분류에 기반한 한국어 형태소 분석 및 품사 태깅", 한글 및 한국어정보처리, 2013

[5] 김영길 외, "지식학습 기반의 다국어 확장이 용이한 관광/국제행사 통역률 90%급 자동 통번역 소프트웨어 원천 기술 개발", 지식경제 기술혁신사업 연차보고서(1차년도), 2013

[6] G. Kikui 외 "Creating corpora for speech-to-speech translation", EUROSPEECH, 2003

[7] 이담허 외 "여행분야 대화체 한일 통계기반 번역시스템 구현", 한국컴퓨터종합학술대회, 2013

[8] S. J. Lee 외 "Intra- and Inter-Frame Features for Automatic Speech Recognition," ETRI Journal, accepted for publication

[9] 박전규 외, "스마트TV 음성인식 인터페이스의 설계 및 구현", 한국음성학회 봄학술대회, pp. 184- 185, 2012.

[10] 박전규 외, "스마트TV를 위한 음성인식 서비스 시스템의 구현", 대한전자공학회 하계학술대회, pp. 1856-1857, 2013

정 호 영

1993년 2월 경북대학교 전자공학과 (학사)
1995년 2월 KAIST 전기및전자공학과 (석사)
1999년 8월 KAIST 전기및전자공학과 (박사)
1999년 8월~현재 한국전자통신연구원 책임연구원

〈관심분야〉
음성인식, 잡음처리, 머신러닝

전 형 배

1999년 2월 연세대학교 전자공학과 (학사)
2001년 2월 KAIST 전기및전자공학과 (석사)
2001년 3월~현재 한국전자통신연구원 선임연구원

〈관심분야〉
음성인식, 언어모델

박 기 영

1997년 2월 KAIST 전기및전자공학과 (학사)
1999년 2월 KAIST 전기및전자공학과 (석사)
2003년 8월 KAIST 전기및전자공학과 (박사)
2003년 9월~2005년 9월 삼성종합기술원 책임연구원
2005년 9월~현재 한국전자통신연구원 선임연구원

〈관심분야〉
음성인식, 신호처리, 기계학습

박 전 규

1987년 2월 한국외국어대학교 전산과 (학사)
1989년 2월 한국외국어대학교 전산과 (석사)
2010년 2월 배재대학교 정보통신공학과 (박사)
1991년 1월~1999년 12월 한국전자통신연구원 선임연구원
2000년 1월~2001년 1월 L&H Korea 책임연구원
2001년 3월~2002년 4월 Carnegie Mellon University 객원연구원
2002년 4월~2004년 4월 동아시테크(주) 이사/기술연구소장
2004년 4월~현재 한국전자통신연구원 책임연구원

〈관심분야〉
음성언어처리, 자연어처리

김 운

2003년 2월 충남대학교 컴퓨터과학과 (석사)
2007년 2월 충남대학교 컴퓨터과학과 (박사)
2007년 10월~현재 한국전자통신연구원 선임연구원

〈관심분야〉
자동번역, 정보검색, 텍스트마이닝

윤 승

2001년 8월 연세대학교 대학원 국어정보학 협동과정 (석사)
2009년 8월 UST 컴퓨터 소프트웨어 및 공학 (박사수료)
2001년 11월~2006년 11월 한국전자통신연구원 연구원
2011년 6월~현재 한국전자통신연구원 선임연구원

〈관심분야〉
자동통역, 음성인식, 대화처리, 음성인터페이스

이 윤 근

1986년 2월 서울대학교 제어계측공학과 (학사)
1988년 2월 KAIST 전기및전자공학과 (석사)
1998년 8월 KAIST 전기및전자공학과 (박사)
1986년 1월~2000년 3월 LG전자기술원 책임연구원
2000년 4월~2004년 12월 ㈜보이스웨어 연구소장
2005년 3월~현재 한국전자통신연구원 실장/센터장

〈관심분야〉
음성인식, 음성합성, 자동통역, 대화인터페이스