음성인식 연구의 국내외 연구현황과 전망

정 현 열*
*영남대학교 정보통신공학과
E-mail : chy@speech.yeungnam.ac.kr

요 약

본 논문에서는 음성인식기술이 어떻게 변천되어 왔는가를 살펴보고, 음성인식 연구에 관한 최근의 국내외 연구동향과 앞으로의 전망에 관하여 논하고자 한다.

국외의 경우 국가적 차원에서 대규모 프로젝트를 중심으로 연구가 진행되어 음성인식 기술이 크게 발전하여 현재 일부 실용화 시스템이 개발되어 사용되고 있다. 국내의 경우 1980년대부터 비교적 활발한 연구가 이루어져 최근 몇 년간 많은 발전을 가져왔다. 최근에는 대어휘 연속음성인식에서도 신뢰할 만한 결과가 많이 보고되고 있으며, 음성인식 기술 뿐만 아니라 멀티미디어 기술을 이용한 다양한 휴먼 인터페이스를 제공하는 보다 편리한 휴대용 단말기에 관한 연구도 활발해지고 있다.

I. 서 론

음성은 인간이 가지고 있는 기본적인 능력 중에서 가장 중요한 것 중 하나로서 우리가 속박감을 거의 느끼지 않고 자유롭게 구사할 수 있는 가장 자연스럽고 효과적인 정보교류의 수단이라 말 할 수 있다. 또 음성에 의해 표현되는 말은 인간과 인간사이의 의사소통의 수단으로서 뿐만 아니라 논리적으로 사물을 생각하는 경우에 있어서도 중요한 역할을 한다. 이 음성이 인간과 기계와의 통신, 즉, 정보의 교환수단으로도 사용되어 오고 있다.

최근 음성과 자연언어의 기본적인 성질의 이해에 관한 관심도 높아지고 있고 각종 미디어의 발달, 초고속 정보통신망의 구축과 더불어 멀티미디어 통신을 통한 통신판매, 물류처리, 제품홍보등이 폭증하고 있으며 지방자치 시대의 도래와 더불어 관공서의 대민 서비스의 질에 관한 관심도 점점 높아져가고 있다. 이와 더불어 개인용 컴퓨터의 보급의 가속화, 컴퓨터에 의한 신호처리기술과 정보처리기술의 급속한 발전과 더불어 음성을 통한 인간과 기계와의 직접적인 커뮤니케이션을 위한 Man-machine Interface의 중요성도 강조되고 있다. 또 인간과 기계사이 뿐만 아니라 인간과 인간사이에 기계를 넣어 통역을 자동적으로 하고자 하는 연구도 활발히 진행되고 있다.

1960년대부터 음성의 발성과 이해에 관해 많은 기초적 연구가 수행되어온 이래 기계에 의한 연속음성인식, 합성에는 아직 많은 과제가 남아있지만 최근 30-40 여년간 연구결과로 고립단어 인식에 있어서는 많은 발전이 있어 미국, 유럽 일본 등에서는 상용제품도 출현하고 있다 이들 인식시스템의 대부분은 고립단어, 또는 한정된 태스크 범주의 연속음성인식시스템이지만 잡음환경하에서도 95%이상의 인식률을 가진 것이 많다. 인식시스템의 경우, 성능이 향상하는 것에 비례하여 응용분야도 복잡화 다양화되어가고 있다. 예를 들면 각종 자료의 수정 및 관리, 철도 또는 항공편 안내 및 예약, Dictati System, 통역전화, 자동통역시스템, 여행정보안내 시스템 관광안내 시스템 등을 개발하여 상품화하고 있으며 국내에서도 음성구동 퍼스날 컴퓨터, 증권정보안내 시스템이 개발되어 상용화가 진행중에 있고, 미국, 일본 등과 나란히 자동통역시스템 개발사업에도 참여하고 있다[1-3][6]. 또 음성 다이얼링 휴대폰도 개발되어 이용되고 있는 등 그 응용 범위는 광범하다[5].

본 고에서는 이와 같은 다양한 응용분야를 가지고 있는 음성인식 기술에 관한 국내외 연구 동향을 지금까지의 연구현황과 앞으로의 전망으로 나누어 기술하고자 한다.

Ⅱ. 음성인식 기술

2.1 음성인식 기술 개요

음성인식에 관한 연구는 약 40여년간의 역사를 가지고 있으며 그동안 많은 변화를 거듭해 왔다. 현재까지 개발된 여러가지 음성인식 수법 중 선형예측분석법(Linea Predictive Analysis), ⊿Cepstrum, DP(Dyn Programming), HMM(Hidden Markov Model), 통계적 언어모델 등의 기술은 통계적 처리가 중심이 되고 있는 최근의 음성인식 기술의 흐름 속에서 계속 살아 남아있지만 과거에 활발하게 연구된 기술 중에는 통계적 처리에 잘 맞지 않는다든지 하여 존재의미를 잃어버린 기술도 많다. 음성인식기술은 아직 해결하지 않으면 안되는 과제도 많지만 금후 Multimedia 환경 구축에도 여러 형태로 사용될 것으로 기대된다.

2.2. 현재의 음성인식 기술

현재의 음성인식 시스템의 전형적 구성형태를 그림 1에 보인다. 대어휘 연속음성인식에서 Bottom-up 처리를 하면 인식결과의 후보의 수가 폭발적으로 증가해 버린다. 이 때문에 인식대상 태스크에 적합한 언어모델(단어사전, 의미정보, 구문정보, 문맥 정보 등)을 Top-down법을 이용하여 문장가설을 생성(예측)하여 그것을 음소계열로 나타낸 것을 순서대로 입력음성과 조합하여 검증해 가는 방법을 취한다.

입력음성은 음소 조합 전에 5~10ms마다 Spectra Parameter Vector로 변환(特徵分析)하여 두는데 현재까지 개발된 거의 모든 시스템에서 켑스트럼과 그 시간적 변화의 미분계수인 △켑스트럼이 사용되고 있으며 인식의 기본단위인 음소의 음향모델로서는 HMM을 사용하고 있다. 단어사전은 단어의 발성을 음소모델의 계열로 나타내는 것인데 조음결합 때문에 각 음소의 발음은 전후 음소의 영향을 받아 변형된다. 이 때문에 각 음소별로 전후 음소에 의존한 복수의 음소환경의존모델을 이용한다.

언어모델로서는 여러 가지 방법이 연구되고 있으나 가장 많이 이용되고 있는 것으로는 통계적 언어모델이다. 구체적으로는 Bigram, Trigram 등 단어의 연쇄확률이 이용된다. 그 밖의 언어모델로서는 문맥자유문법, 유한상태 네트워크 문법 등이 이용되고있다. 이때 모든 음향적 언어적 제약을 만족하는 가장 가능성이 높은 문장을 탐색하는 알고리즘에 관한 연구도 매우 중요한데 프레임 동기형 빔 탐색, A* 탐색 등이 이용되고 있다.

복수의 지식원을 통합하는 방법으로서 N-best 탐색법도 이상적 방법의 하나로서 널리 이용되고있다. 이것은 우선 간단한 음향모델과 언어모델을 이용해서 N개까지의 인식결과후보를 선택한 후 다음 정도가 높은 모델을 이용해 이들 후보의 순위를 재평가함으로 인식성능을 향상시키는 방법이다[4].

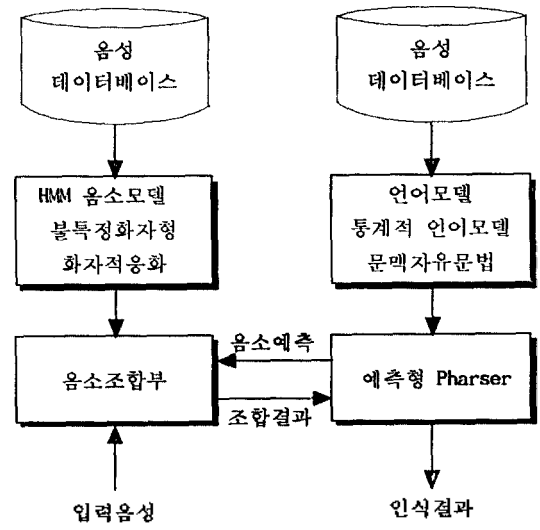

그림 1. 현재의 음성인식시스템의 전형적 구성.

Ⅲ. 현재까지 음성인식기술의 변천

3.1 특징분석

음성인식연구는 1950년대부터 현재까지 세계의 많은 연구자에 의해 40년 이상에 걸쳐 연구되어 왔다. 초기의 음성인식에서는 한 사람의 화자가 또박또박 발성한 한 자리 숫자, 단음절 등을 인식의 대상으로 하고 있었기 때문에 음성의 특징분석으로는 대역필터뱅크(Band Pass Filter Bank)를 이용하는 것이 주류였으나 음운성에 밀접한 관계가 있는 스펙트럼 포락(ENVELOP)을 정도높게 표현하는 것이 어려운 문제였다. 1960년도 후반에 들어 Maximum Likelihood(ML) 추정법이 Itakura에 의해 제안되었는데 이것은 스펙트럼 포락을 전극형모델(All pol model)로 표현하여 최적 모델파라미터를 비교적 적은 계산량으로 안정하게 추출(ML법의 의미)하는 우수한 성질을 갖고 있다. 같은 뜻이지만 시간영역에서는 파형의 선형예측모델로 정식화되므로 선형예측분석(Linea Predictive Coding)법으로도 불려지고 있다.

같은 시기(1960년대)에 스펙트럼 포락을 표현하는 우수한 방법으로서 켑스트럼(Cepstrum)법이 제안되었다. 켑스트럼은 대수스펙트럼의 역퓨리어 변환으로 정의되어 컨버류션(Convolution)관계에 있는 현상을 분리하여 표현할 수 있는 성질이 있고 음성의 스펙트럼 포락과 피치

(Pitch: 기본주파수)를 분리 추출할 수 있는 방법으로서 주목되었었다. 이 성질은 선형예측분석법에 기반을 둔 피치 추출법 즉, 변형상관법의 제안에도 영향을 끼쳤다고 생각된다.

선형예측분석법은 그 후 우수한 음성의 분석, 합성기술로서 PARCOR (Partial Auto-Correlation)법, LSP(Line Spectrum Pair)법 등으로 발전하여 현재의 음성부호화에도 기본기술로 이용되고 있다. 이와 같은 경향은 음성인식에도 영향을 끼쳐 1970년대부터 선형예측분석법에 기초를 둔 음향모델이 전적으로 사용되게 되었다.

통계적 패턴인식이론을 이용하는 경우 켑스트럼은 선형예측분석으로부터 얻어진 파라미터보다 다음과 같은 점에서 우수한 성질을 갖고 있다.

1) 파라미터에 의해 표현되는 스펙트럼의 포락이 부드럽고 안정한 성질을 갖고있다.
2) 파라미터의 단순한 유크리드거리가 대수 스펙트럼 포락의 거리에 일치한다.
3) 파라미터가 직교화되어 있어 파라미터간 상관이 적다.
4) 파라미터의 분포가 경험적으로 정규분포에 가깝다.

켑스트럼 추출에 필요한 계산량은 선형예측분석보다 약간 많지만 FFT를 이용하면 별 문제없다. 그 후 1970년대 중반에 Bell연구소의 Atal에 의해 선형예측계수로부터 전극형 스펙트럼 포락의 켑스트럼을 직접적으로 구하는 방법이 제안되어 계산량의 문제는 완전히 해결되었다. 또 1970년대말 경에는 청각에 있어서 중요한 음성 스펙트럼의 동적 성질을 화자인식에 이용하는 방법으로서 켑스트럼의 시계열(50~90ms정도)을 10ms정도의 통상의 음성분석frame마다 다항식 전개하여 그 전개관계를 특징파라미터(△켑스트럼 또는 △△켑스트럼)로하여 원 켑스트럼과 조합하여 사용하는 방법이 개발되었다. 그후 1980년도 중반에 이것을 음성인식에도 이용하는 방법이 제안되었다. 이 방법은 음성의 개인차, 잡음, 왜곡 등에 강하고 음성인식의 정도(精度)를 크게 개선하는 효과가 입증되어 그때까지의 선형예측분석에 기초를 둔 방법에 대신하여 세계의 여러 음성인식시스템에 이 방법이 사용되게 되었다. 이렇게 된 배경에는 통계적인 방법이 음성인식의 중심이 되었기 때문에 상기의 켑스트럼이 갖는 우수한 성질이 보다 잘 나타날 수 있었기 때문이다[9].

3.2 DTW와 HMM

1970년대 초에 일본에 있어서의 중요한 연구성과의 하나로 Sakoe에 의한 동적계획법(Dynamic Programming: DP)을 이용한 시간 축 정합법(미국: Dynamic Time Warping: DTW, 일본 DP matching법)이 있다. 이것은 음성의 시간축의 신축에 대처하면서 2개의 패턴의 유사도(거리)를 계산하는 극히 효율적인 방법이다. 이즈음 소련(현 우크라이나)에도 연구되고 있었다는 것이 후에 알려졌다. 이와 같이 음성의 시간신축의 정규화에 동적계획법을 이용하는 방법은 2단 DTW (2 Level DTW)법 또는 그 변형으로서 연속단어음성인식에 대한 단어 열의 정합에도 이용되었으며 현재의 HMM에의한 음성인식에도 이용되고 있다[7-8]

1980년대는 이때까지의 음성의 시간패턴을 직접 정합(매칭)시키는 방법으로부터 HMM을 대표로 하는 통계적 모델화에 기초를 둔 방법에로의 이행으로 특징되어 질 수 있다. HMM에 의한 방법은 IBM, CMU 등의 소수의 연구소에서는 그 이전부터 잘 알려져 있었지만 일반에는 1980년대 중반경 그 구체적인 방법, 이론이 Bell 연구소 연구자들에 의해 널리 발표되어 처음으로 세계의 여러 연구기관에서 이용되게 되었다.

3.3 언어모델

통계적 언어모델은 1950년대에 영국의 University College에서 실시된 음소인식시스템에서 처음으로 영어의 음소연쇄에 관한 통계적 정보가 이용되었다. 그러나 그 후 20년 이상에 걸쳐서 이 방법이 음성인식에 적극적으로 이용되지 못하고 생성규칙에 의한 방법이 전적으로 이용되어 왔었다. 1980년대에 와서 상기의 음향처리에 HMM이 널리 이용되는 것과 보조를 맞춰 언어모델에서도 통계적 방법이 중심적으로 사용되게 되었다. 이 기본이 되는 음성의 발성과 인식모델은 그림 3에 나타낸 바와 같이 통신이론과 대응시켜 정식화할 수 있다.

여기서 화자는 정보원에 대응하는 문장 w를 그 화자의 습성에 따라 발성하고 음성인식시스템은 그 음성으로부터 음향처리(분석)에 의해 음향파라미터 y를 얻는다 이 모델에서는 화자와 음향처리를 하나로 묶어 음향채널로 생각해 통신이론에 있어서의 잡음이 부가된 통신로와 대응시키고 있다. 언어복호부는 y를 기지로 하여 사후확률 $p(w|y)$가 최대로 되는 w를 구하는 것을 목적으로 하지만 이것은 직접적으로 계산할 수 없으므로 Bayes Rule에 따라 변형하여 동시확률 $p(w,y)=p(y|w)p(w)$를 최대화 한다. 이때 $p(y|w)$는 HMM에 의해 주어지고 $p(w)$는 통계적 언어모델에 의해 주어진다[11-13].

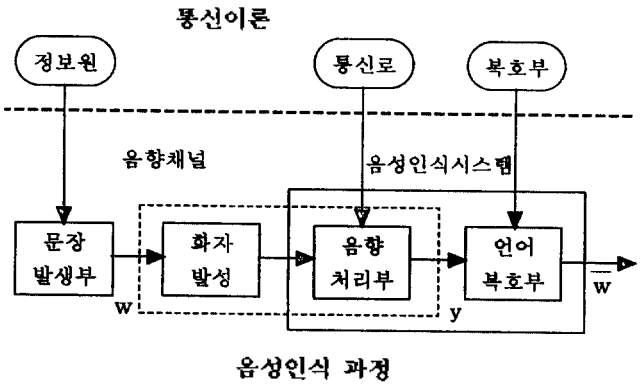

그림3. 현재 음성인식시스템의 기본이 되고 있는 발성과 인식모델

$$p(\overline{w}|y) = \max_{w} \frac{p(y|w)p(w)}{p(y)}$$

1980년대에 통계적언어모델이 특히 미국에서 널리 이용된 직접적 원인은 ARPA(Advanced Research Projec Agency)에 의한 1,000 ~ 20,000 어휘를 대상으로 한 연속음성인식 프로젝트에 의해 막대한 양의 언어 및 음성 데이터베이스가 구축될 수 있게 되면서 부터이다. 이 프로젝트에서는 인식결과의 평가법을 명확히 하여 결과에 대한 Contest를 매년 실시함으로써 치열한 경쟁을 연구에 도입하였고 이 결과 세계의 음성인식기술 수준 향상에 크게 이바지하게 되었다. 유감스럽게도 우리나라에서는 이와 같은 공통의 데이터베이스를 구축하는 움직임이 없었기 때문에 연구가 크게 뒤지고있다.

3.4 Neural Network

1980년대의 뉴럴 네트워크 붐이 한창일 때 음성인식 분야에서도 뉴럴 네트워크를 이용하는 시도가 많이 이루어져 한정된 태스크에 있어서는 높은 인식 성능을 나타내어 주목을 받은 때가 있었다. 그러나, 결국 HMM과 같은 통계적 방법을 명확히 능가하지는 못했다. 또, 연속음성인식과 같은 음향모델과 언어모델을 복잡하게 조합하지 않으면 안되는 경우에서는 비선형적인 처리결과를 종합적인 스코어로 축적시킨다든지 하여 문장 전체로서의 인식성능을 높이는 파라미터의 최적화를 꾀하는 것이 매우 힘들기 때문에 최근 음성인식에 있어서의 뉴럴 네트워크의 연구는 열기가 식어가고 있다. 단, 뉴럴 네트워크의 일종인 학습벡터 양자화(LVQ: Linear Vecto Quantization)의 이론적 발전과 더불어 파생된 오인식별 소화학습(MCE/GPD)법은 금후 학습법의 하나로서 현재 주목되고 있다.

Ⅳ. 국내외 음성인식의 연구현황과 전망

4.1 국외 현황

미국의 경우 1970년대부터 미 국방성의 주도에 의한 ARPA 프로젝트의 일환으로 연속음성인식에 대한 본격적인 연구가 시작되어 진행되던 중 1984년부터 DARPA 프로젝트로 바뀌어 대용량 어휘 음성인식 및 구어체 언어 이해 연구가 진행되었다. 이 연구의 장기적인 목표는 100,000 단어의 어휘를 대상으로 한 연속 음성인식이며 단기적으로는 10,000 단어의 어휘를 사용하여 연속음성을 인식하되 95% 이상의 단어 인식률을 얻는 것을 목표로 하고 있다[10]. 이 계획에는 미국 주요 대학 및 연구기관이 참여하여 낭독체의 자원 관리에 의한 질의 및 명령어나 자연스런 대화체의 항공 여행 정보 안내에 의한 질의 및 명령어를 인식하는 것을 공동 목표로 하여 조직적인 성능 평가를 통해 경쟁적인 기술 개발을 유도해 가고 있다.

또한 음성인식과 더불어 터치스크린, 제스처, 문자인식, 얼굴표정, 눈동자의 움직임 등의 비음성 의사표현 수단을 이용한 멀티미디어 에이전트에 관한 연구도 활발하게 이루어지고 있다[41].

최근의 DARPA 프로젝트는 방송 뉴스 연속음성인식 코퍼스인 Hub4를 대상으로 연구가 진행되고 있다[42]. BBN에서는 BYBLOS 시스템을 이용하여 발음 변화, 액센트, 상호조음과 발성 모드에 따른 자연 발화 음성인식의 인식 성능을 개선하여 Hub4 96데이터에 대하여 인식실험을 실시한 결과 84.1%의 단어인식률을 얻었다. 또, GTN 웹사이트와 음성언어 대화 인터페이스를 이용하여 전화를 통하여 정보를 획득할 수 있는 GTNPhone을 개발하였는데, 이것은 무선, 유선 전화를 사용하여 정보를 검색할 수 있다[31][33]. Cambridge 대학에서는 HMM과 MLP(Multilayer Perceptron) 기반의 ABBOT 시스템을 개발하여 1997 DARPA Hub4 데이터를 대상으로 한 인식실험에서 72.9%의 단어인식률을 얻었다[34]. Dragon사는 Hub4에 맞게 Dragon 시스템을 수정하여 성별 독립, 화자정규화 모델을 사용하여 적응화 뒤 인식률은 78.6%의 단어인식률을 얻었다[35]. SRI에서는 GMS(Gaussian Merging Splitting) 알고리즘으로 학습하여 WSJ94S0중 209문장을 대상으로 인식실험을 수행한 결과 88.3%의 인식률을 얻었다[30].

한편, 일본의 NTT에서는 일본어 방송 뉴스를 대상으로 80.7%의 단어인식률을 얻었다[32].

이외에, CMU에서는 MS-TDNN(Multi-State Time Delay Neural Network) 기반의 전화 코퍼스를 인식대상으로 하는 시스템을 구축하여 305개의 전화를 통해 녹음한 사람 이름에 대하여 평균 97.7%의 인식률을 보였다. 그리고

최근 JRTK(Janus Recognition ToolKit)을 사용하여 회의 대화를 인식의 테스크로 하는 Meeting Browser Interface 시스템을 개발하여 3명의 참가자가 회의 대화를 핀 마이크를 통하여 발성된 음성을 대상으로 테스트를 실시한 결과 적응화후 57.2%의 저조한 인식률을 얻었다. 이것은 입력되는 음성이 단일지향성이 아닌 무지향성 마이크로 채록되기 때문에 채널왜곡이 심하고 실제 회의 대화를 대상으로 한 데이터를 이용하였기 때문이다.

또한 음성인식기능을 보완하기 위한 방법으로 오인식을 대상으로 하는 멀티모달 접근법을 도입하였다. 멀티모드로서는 단어 및 철자의 재발성, 필기 등을 이용하였는데, 입력 희망 양식을 테스크로 한 실험에서 단어 인식률이 78%인 경우 제 수정방법으로 철자 재발성을 이용한 경우가 93%로 가장 높았다[23][24][28].

Texas Instruments에서는 여러 가지 다양한 연구가 진행되고 있다. 영어뿐만 아니라 다른 언어를 대상으로 확장하여, 일본어를 사용하여 WWW 브라우저를 제어하는 SAM(Speech -Aware Multimedia)을 개발하여 평균 91.5%의 문장인식률을 얻었다. 또한 전화채널을 통한 10연속 전화번호 인식에서 화자적응화후 99.0%의 단어인식률과 94.3%의 문장인식률을 보였다[15][18].

BBN Systems and Technologies에서는 WSJ(Wall Stree Journal) 코퍼스 중 20명의 화자에 대하여 SAT(Speake Adaptive Training)을 이용한 인식실험 결과 적응화 93.53%, 적응화후 95.18%의 단어인식률을 나타내었다. 또, 연속음성의 효과적 인식을 위해 의미론적 파싱, 의미론적 분류, 화법 모델링의 세단계를 거쳐 ARPA Air Travel Information System(ATIS) 태스크에 대하여 시험결과 90.3%의 문장인식률을 얻었다[14][17].

Bell 연구소에서는 inter-word 문맥종속 모델을 MCE (Minimum Classification Error) 학습을 사용하여 2,986 전화번호를 대상으로 하는 화자독립 인식실험을 수행하여 90.9%의 인식률을 얻었다[22].

MIT에서는 GALAXY 시스템에 기반한 전자적 자동차 분류 광고 데이터베이스에 접근하여 정보를 제공해주는 대화 시스템인 WHEEL을 개발하여 5,000 종류의 자동차의 데이터베이스를 검색하기 위한 1,200개의 발성을 대상으로 한 성능평가 결과 76.3%의 인식률을 얻었다[25] 역시 GALAXY에 기반한 식당 안내 시스템인 DINEX를 개발하여 보스턴 시내 450개의 식당을 대상으로 하여 실험한 결과 약 72%의 인식률을 얻었다[26].

일본의 경우 1986년이래 15년간의 장기 계획으로 자동통역전화개발계획을 추진해 오고 있으며, 1987년에는 국가 주도에 의한 인간과 기계화의 구어체대화를 목표로 하는 "Advanced Man-Machine Interface through Spok Language" 계획이 시작되어 대화체 언어이해 및 소음 환경에서의 음성인식에 관한 연구가 진행되어 많은 결과를 도출하였다.

NTT에서는 음성에 의한 홈뱅킹 시스템을 개발하여 전화를 통하여 7안속 숫자, 은행 이름, 돈 액수 등을 대상으로 하여 약 85%의 인식률을 얻었다[29].

동경공대에서는 문맥종속 음소모델과 단어 trigram을 이용한 데어터 연속음성인식 시스템을 개발하여 10명의 화자가 발성한 일본어 경제신문 내용을 대상으로 인식실험을 수행하였다. 언어의 복잡도가 평균 72인 경우 평균 89.9%의 문장인식률을 나타내었다[19].

이외에 음성인식 기술을 응용한 시스템으로서는 토요하시 과학기술대학의 관광안내 시스템, TOSHIBA의 음성인식 자동판매기, NTT의 주소입력시스템, ATR(ATR Interpreting Telecommunications Research Laboratori 성번역통신 연구소의 ASURA 등이 있다[43].

유럽의 경우 1983년부터 ESPRIT 프로젝트를 중심으로 하여 현재까지 약 40여개의 프로젝트를 수행하고 있으며, 최근에는 대화체 음성인식과 사회적 요구에 중심을 둔 새로운 프로젝트들이 진행중에 있다. 대표적인 연구로서는 조음처리에서 조음적-음향학적 상관관계, 화자특성의 분석과 합성 등에 대한 기본적인 연구와 더불이 음성처리를 위한 보다 향상된 알고리즘과 구조에 관한 연구, 음성의 지능적이고 자식적 인식, Hybrid 시스템을 위한 음성인식 알고리즘, 다중 언어 음성입-출력, Human-Machine 인터페이스에서 음성의 효과적 이용을 위한 연구, 유럽 언어뜰의 언어학적 분석, 다중 인어 음성-텍스트와 텍스트-음성 변환 시스템, 대화 시스템 등 광범위한 연구가 진행되고 있다[37].

이 중에서도 프랑스의 LIMSI 시스템은 연속 혼합 밀도, tied-state cross-word 문맥 의존 HMM을 이용하 Hub4E 데이터에 대해 단어인식률 81.5%를 얻고 있다. 그리고 미국의 Texas Instruments와 함께 연속음성인식시스템을 구성하여 241문장, 복잡도가 31인 테스크를 대상으로 평균 99.31% 단어인식률을 얻었다[16][36].

독일에서는 German VERBMOBILE 프로젝트(영역: 약속 스케줄)에서 구문론적, 운율적 경계에 대한 레이블링을 도입하여 122개의 문장에 대하여 96%이상의 단어인식률을 달성하였다.

4.2 국내 현황

국내에서도 1980년도에 들면서부터 본격적인 음성인식에 관한 연구가 이루어져 오고 있다. 개발된 시스템으로는 한국전자통신 연구소의 자동통역시스템, 한국통신의 증권정보 안내시스템, 삼성전자의 음성구동 퍼스날 컴퓨

터, 음성구동 셀룰러폰(삼성, LG), 음성메모장치(공성통신)등이 있으며 현재 성능개선 또는 상용화중에 있다. 또 음성에 의한 로보트 제어에 관한 연구, 음성에 의한 자동항법 장치 등에 관한 연구도 활발히 진행되고 있다. 이하 몇몇 연구기관들의 연구에를 구체적으로 나열하기로 한다.

한국과학기술원 음성언어연구실에서는 Triphone 모델을 인식단위로 한 화자독립 연속음성인식에서 3,000단어 규모의 연속음성인식 시스템을 개발하여 단어인식률 92.19%, 문장인식률 67.8%를 달성하고 있다[20].

한국전자통신연구원에서는 1995년부터 Human-Computer Interface를 위한 음성 입/출력 처리에 관한 연구로 5,500 단어 규모의 연속음성 번역 시스템(한국어-영어 한국어-일본어)을 개발하고 있으며, 현재 한-영 번역 시스템에서 평균 79.1%의 변환율을 얻고 있다. 그리고, 멀티모달 휴먼인터페이스에 관한 연구도 진행되고 있다[21].

한국통신에서는 증권정보 안내시스템과 함께 호텔예약 테스크에 대한 번역시스템을 개발하여 총 30,204 어절을 대상으로 한 실험에서 92.3%의 변환률을 얻고 있으며 성능개선을 거듭하고 있다[27].

영남대 음성언어처리 연구실에서는 연속음성인식 기술을 이용한 대화형식의 항공편 예약 시스템을 개발하여 어휘 수 346에, 평가용 100문, perplexity 43.2에 대 평균 문장인식률 68.3%, 단어인식률 86.4%를 보이고 있으며, PC 윈도우 환경에서 동작하는 음성인식기능을 가진 주소 입력검색 시스템을 개발하여 on-line 테스트에서 평균 90.7%의 연결단어 인식률과 97.7%의 단어인식률을 얻었다[38][39]. 또한 음성인식 기능을 이용한 지도검색 시스템을 개발 중에 있다[40].

4.3 음성인식 기술의 21세기에의 전망

최근 음성인식기술은 미국을 중심으로 구체적인 응용분야가 개척되어오고 있고 멀티모드/멀티미디어 환경 속에서의 다른 미디어와 통합에 관한 연구가 진행되고 있다. 향후 이러한 멀티미디어와 결합되는 연구가 더욱 활발하게 진행될 것으로 기대된다. 이러한 멀티모드/멀티미디어 기술의 활용분야로는 각종 멀티미디어 정보기기의 입출력 인터페이스, 카 네비게이션 시스템 개발, 시각장애자를 위한 서비스 시스템 개발, 대화형 자판기, 대화형 Robot, 3차원 컴퓨터 시스템 개발, 제품의 검사, 멀티모드 의료 서비스, 각종 멀티모달 데이터 베이스 검색 멀티모드형 인터넷 검색기, 홈쇼핑, 자동 예약/문의 시스템, 음성 입출력 PC, 전자 메일 시스템 개발, 멀티모드형 자동항법 장치 개발, KIOSK 개발 등 그 분야는 이루 헤아릴 수 없다.

이와 같은 응용연구와 더불어 자연어 처리기술을 적극적으로 이용하는 자연발화 대화체 연속음성인식에 관한 연구가 더욱 활발하게 진행될 것 생각된다. 이와 더불어 각국간의 자동통역전화에 관한 연구도 가속화될 것으로 보인다.

음성인식 전반으로서는 현재의 통계적 방법을 기반으로 실제의 대량의 음성 데이터에 기초를 둔 일상 언어의 언어모델(규칙)을 구축하는 것, 다수화자의 음성데이터에 기저하여 개인차의 모델을 구축하여 이에 의한 다수 화자의 음성에의 적응화 알고리즘을 개발하는 것, 여러 종류의 잡음, 왜곡에 자동적으로 적응되는 방법을 확립하는 것 등이 중요한 기술적 과제로 될 것이다.

국내적으로는 하루빨리 대규모 한국어 음성데이터베이스가 구축되어 많은 음성연구자들이 공동으로 이용하여 서로의 연구결과를 평가하고 그 결과를 공유할 수 있는 기반이 조성되어야 할 것으로 생각된다.

Ⅳ. 결론

음성인식 기술에 관한 국내외 연구동향과 앞으로의 전망에 대하여 고찰하였다.

지난 40여년간의 고립단어, 연속음성인식에 관한 기술을 기반으로 최근 미국을 중심으로 구체적인 응용 분야가 널리 개척되어오고 있고 현재 멀티모드/멀티미디어 환경 속에서 다른 미디어와 결합하고 있으며, 이러한 멀티모드와 결합된 연구가 더욱 활발하게 진행될 것으로 기대된다. 이러한 멀티모드/멀티미디어 기술의 활용분야로는 각종 멀티미디어 정보기기의 입출력 인터페이스, 카 네비게이션 시스템, 시각 장애자를 위한 서비스 시스템, 대화형 자판기, 대화형 Robot, 멀티모드형 인터넷 검색기, 홈쇼핑 등 무수히 많으며 이와 같은 분야에의 상품화 연구가 활발할 것으로 생각된다.

이와 같은 응용연구와 더불어 자연어 처리기술을 적극적으로 이용하는 자연발화 대화체 연속음성인식에 관한 연구가 더욱 활발하게 진행될 것 생각된다. 이와 더불어 각국간의 자동통역전화에 관한 연구도 가속화될 것으로 보인다.

음성인식 전반으로서는 현재의 통계적 방법을 기반으로 실제의 대량의 음성 데이터에 기초를 둔 일상 언어의 언어모델(규칙)을 구축하는 것, 다수화자의 음성데이터에 기저하여 개인차의 모델을 구축하여 이에 의한 다수 화자의 음성에의 적응화 알고리즘을 개발하는 것, 여러 종류의 잡음, 왜곡에 자동적으로 적응되는 방법을 확립하는 것 등이 중요한 기술적 과제로 될 것이다.

참고문헌

1. 안수길, "한국에서의 음성신호처리 기술의 현황과 전망," 제12회 음성통신 및 신호처리 워크샵 논문집 1994.6.
2. 김순협, "음성 인식기술의 현황 및 실용화 전망," 한국음향학회지, 13권, 2호, 1994.
3. 구명완, "음성 인식기술의 현황과 실용화 전망," 한국음향학회 학술발표회 논문집, Vol. 17, No. 1, 1998.7
4. "音聲認識," 日本電子情報通信學會志 Vol.78 No.1 pp.1114-1118, 1995.11.
5. S.H.Choi et al., "Continuous Digit Recognition Rreal-Time Voice Dialing System Using Discr Hidden Markov Models," Proc. 5th WESTPRAC, pp.1027-1031, 1994.
6. M.W. Koo et al., "KT-Stock: a Speaker-Indepen Large Vocabulary Speech Recognition System o The Telephone," Proc. ICSLP, pp.1387-1390, 1994.
7. A. Kai and S. Nakagawa, "A Frame-Synchronou Continuous Speech Recognition Algorithm Using Top-Down Parsing of Context-Free Grammar," Proc. ICSLP 92, pp.119-121, 1992.
8. H.Ney, "Dynamic Programming Parsing for Context F Grammars in Continuous Speech Recognition," IE Trans., Vol.3, No.3, pp.336-340, 1990.
9. L.R.Rabiner, B.H.Jang, "Fundamentals of Sp Recognition," prentice Hall, 1993.
10. D.S.Pallet, "DARPA Resource Management and ATI Benchmark Test Poster Session," Proc. of the DA speech and Natural Language Workshop, pp.49- Feb., 1991.
11. M.Lenning et al., "Flexible Vocabulary Recognit Speech," Proc. of ICSLP 92, pp.93-96, Oct, 1992.
12. G.G.Matison, "Emerging Voice Services in the Ny Network," Proc. of voice Systems Worldwide 1 pp.9-13, Feb. 1992.
13. R. Kompe et al., "Improving Parsing of Sponta Speech with the Help of Prosodic Boundaries," of 1997 IEEE ICASSP, pp.811-814, 1997.
14. Tasos Anastasakos, John McDonough, and Jo Makhoul, "Speaker Adaptive Training: A Maximu Likelihood Approach to Speaker Normalization," of 1997 IEEE ICASSP, pp.1043-1046, 1997.
15. Kazuhiro Kondo and Charles T. Hemphill, "Surf The World Wide Web with Japanese," Proc. of 19 IEEE ICASSP, pp.1151-1154, 1997.
16. Irina Illina, Yifan Gong, "Elimination of Trajectory Folding Phenomenon: HMM, Trajectory Mixture HMM and Mixture Stochastic Trajectory Model," Porc. of 1997 IEEE ICASSP, pp.1395-1398, 1997.
17. Richard Schwarts, Scott Miller, David Stallar John Makhoul, "Hidden Understanding Models f Statistical Sentence Understanding," Proc. o IEEE ICASSP, pp.1479-1482, 1997.
18. Yu-Hung Kao and Lorin Netsch, "Inter-Digit H Connected Digit Recognition Using The Macropho Corpus," Porc. of 1997 IEEE ICASSP, pp.1739 17 1997.
19. Tatsuo Matsuoka et al., "Japanese Large-Vocab Continuous-Speech Recognition Using Business-Newspaper Corpus," Proc. of 1997 I ICASSP, pp.1803-1806, 1997.
20. Seong-Jin Yun, Yung-Hwan Oh, Gyung-Chul Shin "Improved Lexicon Modeling For Continuous Spee Recognition," Proc. of 1997 IEEE ICASS pp.1827-1830, 1997.
21. Nam-Yong Han, Un-Cheon Choi, Young-Jik Lee, "A Implementation of a Partial Parser in The S Language Translator," Proc. of 1998 IEEE ICAS pp.205-208, 1998.5.
22. Malan B. Gandhi and John Jacob, "Natural Numb Recognition Using MCE Trainde Inter-Word Conte Dependent Acoustic Models," Proc. of 1998 IE ICASSP, pp.457-460, 1998.5.
23. Hua Yu, Cortis Clark, Robert Malkin, Alex Wai "Experiments in Automatic Meeting Transcription JRTK," Proc. of 1998 IEEE ICASSP, pp.921-92 1998.5.
24. Hermann Hild, Alex Waibel, "Recognition of Spe Names over The Telephone", Proc. of ICSLP 9 pp.346-349, 1996.10.
25. Helen Meng et al., "WHEELS: A Conversation System in The Automobile Classified Domain", P of ICSLP 96, pp.542-545, 1996.10.
26. Stephanie Seneff and Joseph Polifroni, "A Restaurant Guide Conversational System: Issu Rapid Prototyping for Specialized Domains", Pro ICSLP 96, pp.665-668, 1996.10.
27. Youngkuk Hong, Myoung-Wan Koo, Gijoo Yang, "A Korean Morphological Analyzer For Speech Transla

System," Proc. of ICSLP 96, pp.673-676, 1996.10.

28. Bernhard Suhm, Brad Myers, and Alex Waibe "Interactive Recovery From Speech Recognition Errors in Speech User Interfaces," Proc. of ICSLP 96, pp.865-868, 1996.10.
29. Toshihiro Isobe et al., "Voice-Activated Home B System and Its Field Trial," Proc. of ICSLP pp.1688-1691, 1996.10.
30. Ananth Sankar, "Experiments with a Gaus Merging-Splitting Algorithm for HMM Training Speech Recognition," Proc. of DARPA Broadca News Transcription and Understanding Worksh 1998.2.
31. Daben Liu et al., "Improvements in Spontan Speech Recognition," Proc. of DARPA Broadca News Transcription and Understanding Worksh 1998.2.
32. Sadaoki Furui et al., "Japanese Broadcast Transcription and Topic Detection," Proc. of D Broadcast News Transcription and Understan Workshop, 1998.2.
33. David Stallard, Joshua Bers and Christopher B "The GTNPhone Dialog System," Proc. of DARPA Broadcast News Transcription and Understan Workshop, 1998.2.
34. G.D. Cook, A.J. Robinson, "The 1997 ABBOT Syste for The Transcription Broadcast News," Proc DARPA Broadcast News Transcription an Understanding Workshop, 1998.2.
35. Steven Wegmann et al., "DRAGON System's 1997 Broadcast News Transcription System," Proc. DARPA Broadcast News Transcription an Understanding Workshop, 1998.2.
36. Jean-Luc Gauvain, Lori Lamel and Gilles Adda, LIMSI 1997 Hub-4E Transcription System," Proc. DARPA Broadcast News Transcription an Understanding Workshop, 1998.2.
37. Joseph Mariani and Lori Lamel, "An Overview of Programs Related to Conversational/Inter Systems," Proc. of DARPA Broadcast New Transcription and Understanding Workshop, 1998.2.
38. 오세진, 김범국, 정현열, "연속음성인식 시스템의 능 개선," 1997년도 한국음향학회 학술발표회 논문집, pp.261-265, 1997.11.
39. 황철준, 김득수, 정현열, "On-line 테스트에 의한 멀모달 음성인식 시스템의 성능 평가," 1998년도 한국통신학회 하계종합학술발표회 논문집, pp.1105-1108 1998.7.
40. 김태수, 정현일, "음성을 이용한 수치지도정보 검색시스템의 구현," 1998년도 한국음향학회 학술발표회 논문집, pp.55-58, 1998.3.
41. Alex Waibel, Bernhard Suhm, Minh Tue Vo and Ji Yang, "Multimodal Interfaces for Multim Information Agents," Proc. of ICASSP 97, Vol. pp.167-170, 1997.
42. Proceedings of the DARPA Broadcast New Transcription and Understanding Workshop, 1 1998.2
43. A. kai, S. Nakagawa, " A frame-synchron continuous speech recognition algorithm usi top-down parsing of context-free grammar", ICSLP 92, pp. 257-260, 1992