음성인식기술을 이용한 새로운 서비스

구 명완
한국통신 연구개발원 소프트웨어연구소 음성언어연구팀

New services based on speech recognition technology

Myoung-Wan Koo

Spoken Language Research Team, Software Research Laboratory,
Korea Telecom Research Laboratories

< 요 약 >

본 논문에서는 음성인식기술을 이용한 시스템이 상용화되기 위해서 필요한 기술의 최근 동향과 현재의 기술로 실용화가 이루어지고 있는 서비스등에 대해 알아본다. 최근의 음성인식기술은 실용화를 목표로 음성인식을 위한 기본 유니트 선정, 화자의 음성을 거절하는 기능, 및 실시간 구현 기술에 대한 연구가 활발히 진행되고 있다. 한편 현재의 기술로 가능한 실용서비스로는 전화번호안내, 음성다이얼링 서비스등과 같이 현재 제공되고 서비스의 비용을 절감시키는 것과 교통안내, 날씨안내, 영화관 예약에 음성인식기술을 적용하여 새로운 서비스를 제공하는 것이 있다.

1. 서 론

음성인식연구는 1952년 미국 벨 연구소의 숫자음 인식기인 Audrey에 관한 연구로부터 40년이상 계속 진행되어 왔으나 아직까지 일반적으로 사용될 수준의 실용화된 시스템은 거의 없다. 그러나 최근 몇년간 실용화를 목적으로 한 연구의 괄목할 만한 성장이 있었다. 특히 미국 국방성 산하의 DARPA(defence advanced research project agency)에 의해 RM(naval resource management), ATIS(air travel information system)및 WSJ(wall street journal)에 관한 음성데이타베이스가 확보되면서 여러기관 간의 음성인식 성능 경쟁이 치열해졌다[1]. 일본에서는 자동통역 전화시스템 개발과 전화번호 안내를 위한 연구가 활발히 진행되고 있으며[2] 유럽에서는 필립스, 캠브리지 대학등을 중심으로 음성인식 연구가 진행되고 있다[3].

음성인식기술의 목표는 기계를 음성으로 명령하여서 작동시키는 것이다. 그런데 현재는 음성명령대신 키이보드, 터치패널 및 마우스등의 입력수단을 이용하여 기계를 작동시키기 때문에 음성명령이 다른 입력수단에 비해 경쟁력이 있기 위해서는 모든 사용자와 모든 환경에서 높은 인식률을 유지할 수 있어야 한다. 이러한 음성인식은 패턴인식방식에 근거하며 그림 1 에서와 같은 과정으로 이루어진다. 먼저 음성이 입력되면 음성의 특징이 추출되어 비교기로 간다. 비교기에서는 음성인식을 위한 기본단위와 단어사전을 이용하여 단어단위로 인식하게 된다. 이때 입력된 음성이 단어이면 확인과정을 거쳐 인식을 하게되고 문장인 경우 형식 및 의미해석기에 의해 문법과 업무모델에 따라 문장이 인식되고 확인과정을 거친다.

본 논문에서는 현재 가장 많이 사용되고 있는 인식기술인 HMM(hidden Makov model)을 사용한 인식시스템이 실용화되기 위해 필요한 주요기술을 서술하고 현재 진행되고 있는 응용사례를 살펴본다. 먼저 2 장에서는 음성인식에 사용되는 기본 유니트 선정, 인식된 단어를 확인하는 거절기능 및 음성인식시스템의 실시간 구현에 대해 기술하고 3 장에서는 음성인식기술을 이용한 응용기술을 비용을 절감하는데 사용되는 서비스와 새로운 수입을 창출할 수 있는 서비스로 나누어서 설명한다.

2. 응용 기술

현재 음성인식시스템을 상용화시키기 위해 필요한 대표적인 기술은 기본 인식단위 선정, 인식 거절기능 및 실시간 알고리즘 구현등이 있으며 각 분야의 기술현황은 다음과 같다.

2.1. 기본 인식단위 선정

음성인식시스템의 기본 인식단위는 인식대상 어휘량, 저장용량 및 인식률에 따라 바뀐다. 단어를 기본 인식단위로 선정하면 인식률은 항상이 되나 인식대상 어휘량이 증가함에 따라 저장용량도 증가하게 되고 인식시간도 증가하게 된다. 또한 새로운 어휘가 인식대상 리스트에 포함되게 되면 그 단어에 대한 음성 데이타베이스가 많이 필요하며 새로운 훈련과정도 필요하다. 최근에는 이러한 단점을 보완하기 위해 서브워드를 기본 인식단위로 선정한다. 서브워드는 언어에 따라 음절, 반음절, 유사음소등으로 정의되는데 영어권에서는 유사음소를 주로 사용하고 중국어인 경우 음절을 사용하기도 한다[4].

일반적으로 가장 많이 사용되고 있는 기본 유니트인 유사음소는 언어의 음운현상을 고려하여 음소 갯수보다는 많으나 조음현상을 고려한 문맥 독립 유사음소(context-independent phoneme like unit)와 문맥종속 유사음소(context-dependent phoneme like unit)로 나눌수 있다. 문맥독립 유사음소는 주위의 유사음소에 영향을 받지 않도록 모델링된 음소이며 문맥종속 유사음소는 주위의 문맥에 따라 유사음소의 종류를 달리하여 모델링된 음소이다. 대표적인 문맥 종속 유사음소로 triphone이 있다[4]. 연속음성인식시스템에서는 단어와 단어사이에도 음운변동현상이 생기므로 단어간(inter-word) 조음현상을 고려해야 한다[5][6]. 또한 a, the, in, and 등과 같이 조음현상을 규명하기 어려운 단어인 경우 단어전체를 하나의 기본 유니트로 모델링하는 기능단어의존(function-word-dependent) 유니트를 사용하기도 한다[4]. 그러나 이러한 문맥종속 유니트를 사용할 경우 훈련데이타에 모델링되고자 하는 유니트가 존재해야 하며 갯수도 많아야만 높은 인식률을 얻을 수 있다. 실제로 이러한 조건을 만족하는 훈련데이타를 매 응용시스템마다 구한다는 것은 실용시스템 설계에 문제가 있게 된다. 최근에는 응용시스템 및 훈련데이타에 관계없이 기본 유니트를 선정하고 모델링하는 방법에 대한 연구가 활발히 진행되고 있다[7]. 또한 문맥 정보를 보다 포괄적으로 이용하기 위해 많이 사용되고 있는 음소는 몇개의 음소를 결합시켜 새로운 기본 유니트로 사용하고 그렇지 않을 경우 음소를 기본 유니트로 사용하는 방법도 개발되고 있다[8].

한편 기본 인식 단위가 선정이 되면 훈련데이타를 사용해서 기본 유니트를 모델링하여야 한다. 훈련데이타의 양은 가능한 많아야 하지만 실제로는 많은 데이타를 얻기가 불가능하기 때문에 인식대상 어휘가 가능한 많이 포함된 훈련데이타를 구하여야 한다. 이를 피하기 위해 기본 유니트에 대한 일반적인 모델을 한 후 매 응용시스템에 대해 실제적인 음성데이타를 얻어서 기본 유니트의 모델을 향상시키는 적응 훈련방식이 제안되었다[9].

2.2. 인식 거절기능

인식 거절기능이란 입력된 음성으로선 인식하기 어려운 상태를 나타내는 것으로 상용 시스템이 되기 위해선 필수적이다. 특히 고립단어 인식시스템을 일반 사용자들이 사용할 경우 단어 이외의 음성을 말하는 것은 지극히 당연하다. 그러므로 입력음성과 기준패턴을 비교하여서 가장 유사한 단어를 선택하는 기존 알고리즘을 그대로 사용하면 많은 문제점이 발생한다. 이를 위해 입력음성으로부터 키어워드(keyword)를 찾아내고 찾아낸 키어워드가 제대로 인식되었는 지를 확인하는 작업이 인식 거절기능에 포함된다. 인식 거절기능이 포함된 음성인식 시스템의 평가는 다음과 같이 네가지 종류의 평가항목이 필요하다[10].

(a) 바른 인식(correct acceptance)

입력된 음성의 의미에 맞게 제대로 인식

(b) 바른 거절(correct rejection)

인식대상 단어 이외의 음성이 입력된 경우 인식 거절을 한 경우

(c) 틀린 인식(false acceptance)

키어워드 A를 키어워드 B로 잘못 인식하거나 키어워드가 아닌 음성을 키어워드로 잘못 인식(false alarm)하는 경우

(d) 틀린 거절(false rejection)

키어워드를 검출하지 못하는 경우

(a) + (b) + (c) + (d) = 100%

전기통신망에서의 응용시스템을 개발할 경우 대부분의 입력음성이 키어워드 주변에 여분의 음성이 약간만 존재하므로(c) false alarm에 중점을 두며 연속음성에서 키어워드를 모니터하는 응용시스템에서는 (d) 틀린거절에 중점을 두어 개발하고 있다.

키어워드를 인식하고 인식거절기능을 구현하는 방식은 여러가지 종류가 제안되었다. 일반적으로 키어워드이외의 단어를 filler model 혹은 garbage model이라고 정의하여 모델링하고 기존의 패턴인식 알고리즘을 그대로 사용한다. 이 때 filler model은 그림 2 와 같이 한 단어처럼 모델링하거나 여러개의 대표적인 단어로 나누거나 혹은 음소모델링 및 문법정보를 사용하기도 한다[11][12].

Filler model이 훈련과정에서 완성되면 인식과정에서는 첫번째 후보단어와 두번째 후보단어의 차이가 작으면 인식거절이 되고 차이가 크면 첫번째 후보단어를 인식된 단어로 선정한다. 이때 첫번째 후보단어가 filler일 경우는 입력된 음성에 키어워드가 없다고 가정하고 인식거절을 하게 된다.

2.3. 실시간 처리

음성인식 시스템이 실용화되기 위해서는 저가의 하드웨어를 사용하여 실시간에 동작이 되어야 한다. 그러나 음성인식 소프트웨어는 많은 계산을 필요로 하기때문에 실시간 처리를 위해선 고속의 하드웨어를 필요로 하였으며 고속의 하드웨어는 고가이기 때문에 음성인식 시스템의 가격이 높았다. 최근에는 하드웨어의 성능에 비해 가격이 많이 떨어졌지만 음성인식 소프트웨어를 실시간으로 동작시키기 위해선 특수한 하드웨어가 필요하다는 단점이 있었다.

BBN에서는 대용량 음성언식시스템을 특별한 하드웨어없이 실시간으로 처리할 수 있는 소프트웨어를 개발하였다[13]. 이

소프트웨어는 음성인식시스템에서 가장 많은 시간을 필요로하는 검색부분을 개선한 것으로 알고리즘의 개선만으로도 실시간 처리가 가능하다는 사실을 제시하였다. 제안된 알고 forward-backward 알고리즘으로서 전방향(forward) 검색에는 간단한 정보만을 이용하여 빠른 시간에 검색이 이루어지도록 하고 후방향(backward) 검색에는 상세한 정보를 사용하여 인식성능을 향상시키도록 하는 알고리즘이다[14].

검색알고리즘을 구현하는 방법은 종합처리(integrate approach)방식과 모듈처리(modular approach)방식으로 나누어진다[15]. 종합처리방식에서는 모든 지식정보를 한꺼번에 이용해서 음성을 인식하는 방식이다. 즉, 음성특징, 발음사전, 구문 및 의미정보가 하나의 finite state로 표현된다. 현재 이러한 방식은 소용량 단어 및 단순한 문법으로 구성된 연속음성인식시스템에 사용되며 인식시간이 적게 걸리며 시스템이 단순하다는 장점이 있다. 그러나 모든 지식정보가 한번에 종합화 될 수 없는 경우가 발생될 경우 이러한 방식을 이용하기가 어렵다. 예를 들면 prosody혹은 trigram과 같은 정보는 쉽게 finite state형태로 변형시키기 어렵다. 또한 단어의 갯수가 증가되면 그에 따라 finite state가 복잡해지므로 실제 대용량 음성인식 시스템에는 적합하지 않다.

반면 모듈처리방식은 음성인식단계를 여러모듈로 나누고 매 모듈에서는 모듈정보를 활용하여 최종 결과를 찾는 방식이다. 예를 들면 음성특징을 이용하여 단어를 인식하고 그 다음에 구문정보를 이용하여 문장을 일 차로 인식하며 최종적으로 의미정보를 활용하여 인식된 문장을 결정한다. 그러므로 언어처리, 음성처리 등의 알고리즘이 독립적으로 개발될 수 있으며 단어가 많아지거나 문법이 복잡하여도 모듈단위로 계산이 이루어지므로 대용량 연속음성 인식시스템에 적합하다. 그러나 매 모듈에서의 정보는 다른 모듈에서의 정보와 독립적으로 사용되므로 한 모듈에서 정보를 잘못 사용하면 다음 모듈에서는 복구할 수 없다. 그러므로 매 모듈에서 정보를 잘 사용하여야 한다.

현재까지의 개발된 대표적인 검색 알고리즘은 다음과 같다.

(a) Frame-synchronous beam 검색

이 검색방식은 매 검색이 진행되는 동안 매 시간마다 나타나는 모든 검색영역을 유지시키는 beadth-first 검색 알고리즘으로서 모든 지식을 finite-state network로 표현하며 network상 가장 적당한 경로를 선택하는 것이다. 검색시간을 줄이기 위해 beam width 혹은 beam thresh-old값을 정해 검색영역을 줄이거나 발음사전을 tree 구조로 구성하여 중복된 계산을 피하기도 한다[16].

(b) Stack decoding and A* heuristic 검색

이 검색방법은 매 시간마다 지엽적인 언어정보를 사용하여 검색하는 best-first 검색알고리즘으로서 매 시간마다 stack을 사용하고 stack상의 최적리스트의 검색영역 models)을 자연스럽게 검색에 사용될 수 있다는 것이다.

(c) Multi-pass decision 검색

이 방식은 첫 검색으로 간단한 정보를 사용하고 두번째 검색으로 자세한 정보를 사용하여 최종적인 결과를 얻는 검색 알고리즘으로서 일반적으로 N개의 적정문장을 찾을 수 있다. N개의 적정문장을 찾을 수 있는 알고리즘으로 sentence-dependent, lattice 및word-dependent 알고리즘이 있으며 word-dependent 알고리즘이 널리 사용되고 있다[17]. 특히 이 알고리즘은 그림 1 에서와 같이 언어정보를 순차적으로 적용할 수 있기때문에 최근에 많이 연구되고 있다.

3. 응용 서비스

전기통신망에서 음성인식기술을 이용한 서비스는 크게 두종류로 나눌 수 있다[18].

첫번째는 비용절감효과를 노리는 서비스이다. 이러한 서비스의 특징은 사람이 해 왔던 일을 음성인식 시스템이 대체하도록 하는 것이다. 그러므로 음성인식 시스템의 성능이 좋아야 한다. 반면 사람대신에 기계가 일을 하므로 비용 절감효과는 있으나 사용자들의 입장에서는 서비스의 질이 떨어졌다고 생각할 수 있다. 대표적인 서비스들은 다음과 같다.

1) 안내양 서비스의 자동화

AT&T에서는 VRCP(voice recognition call processing)를 제공하여 안내양이 제공하던 서비스 일부를 자동화하고 있으며 Bell Northern회사에서는 AABS(automated alternative billing service)를 개발하여 수신자 요금부담 서비스를 안내양없이 자동화하고 있다.

2) 전화번호 안내

Nynex와 Bell Northern회사에서는 도시이름에 대한 음성을 인식하여 해당도시를 담당하고 있는 안내양으로 연결해 주는 서비스를 시험적으로 운용하고 있다.

3) 음성다이얼링 서비스

사람이름 혹은 번호를 음성으로 말을 하면 자동으로 전화가 연결되도록 하는 서비스이다. Nynex에서 시험서비스 중에 있으며 한국통신에서도 개발 중에 있다.

두번째는 새로운 수익을 창출해내는 서비스이다. 이러한 서비스의 특징은 안내양을 사용하여서 서비스해주기에는 비용이 많이 들기때문에 이전에는 서비스가 제공되지 못했다는 것이다. 그러므로 사용자는 이전에 제공받지 못한 정보를 얻을 수 있으므로 음성인식의 성능이 떨어진다고 해도 어느정도 서비스의 질에 대한 불평은 적을 수 있다. 대표적인 서비스들은

다음과 같다.

1) 음성 은행 서비스

NTT ANSER 시스템은 음성인식기술을 이용하여 계좌에 관한 정보를 얻을 수 있다. ANSER 시스템은 전자식 전화기를 사용하지 않는 사람들을 위해 음성인식기술을 제공한다.

2) 정보검색시스템

Northen Telecom이 제공하는 증권정보 안내시스템이 있다. 이 시스템은 음성으로 상장회사이름을 입력하면 인식하여 해당회사의 주식정보를 제공하는 서비스를 수행한다. 한국통신에서도 개발되어 시험 중에 있다[19].

3) 전화번호 안내 및 자동 다이얼링

NYNEX 및 AT&T가 제공해주는 서비스로서 전화번호 안내가 제공되는 동시에 간단한 제어명령을 인식하여 자동적으로 안내된 전화번호로 다이얼링되도록 해준다.

4) 가입자 정보 안내

전화번호를 음성으로 말할 하면 가입자의 이름과 주소를 알려주는 서비스이며 NYNEX, Bellcore 및 Ameritech 회사에서 제공한다.

5) 정보서비스

교통안내, 날씨안내, 영화관 예약등과 같은 서비스는 음성 인식기술을 이용하여 제공한다.

4. 결 론

이 논문에서는 음성인식기술을 이용한 시스템을 실용화시키기 위해 필요한 기술 중 음성인식을 위한 기본 유니트 선정, 인식 거절기능 및 실시간 구현에 대해 현 기술의 현황에 대해 기술하였으며 음성인식기술을 이용한 응용서비스에 대하여 알아 보았다. 최근의 급격한 음성인식기술의 성장으로 외국에서는 전기통신망을 통한 음성인식 응용서비스들이 점차로 나타나고 있는 추세이며 상용화의 장애가 되는 분야의 연구도 활성화되고 있다. 국내에서는 한국통신을 중심으로 상용서비스의 기틀을 다지고 있으며 범용 PC상에 간단한 명령을 음성으로 입력할 수 있는 소프트웨어가 일부 기업에서 보급되고 있다. 앞으로 연구결과가 더욱 진척되면 일반사용자들이 보다 손쉽게 음성인식기술을 이용한 응용서비스를 접할 수 있게 될 것이다.

참 고 문 헌

[1] D. Pallet, et al., "DARPA February 1992 ATIS Benchmark test results," *Proc. of the DARPA Speech and Natural Language Workshop*, pp. 15-27, 1992.

[2] S. Sagayama, et al., "ATREUS : a speech recognition front-end for a speech translation system," *Proc. of Eurospeech 93*, pp. 1287-1290, 1993.

[3] J. Peckham, et al., "A new generation of spoken dialogue systems : results and lessons from the SUNDIAL project," Proc. of Eurospeech 93, pp.33-42, 1993.

[4] K. F. Lee, "*Automatic Speech Recognition - The development of the SPHINX-System,*" Kluwer Academic Publishers, Boston, 1989.

[5] E. P. Giachin, et al., "On the use of inter-word context-dependent units for word juncture modeling," *Computer Speech and Language*, pp.197-213, 1992.

[6] C. H. Lee, et al. , "Acoustic modeling for large vocabulary speech recogniton," *Computer Speech and Language*, pp.127-165, 1990.

[7] M. W. Koo, " KARS:Speaker-independent, vocabulary -independent speech recognition system," *Proceeding of EURO SPEECH 93*, pp.1837-1841, 1993.

[8] T. Matsumura, "Toward non-uniform unit HMMs for speech recognition," *Proceeding of IVTTA 94*, pp.89-92, 1994.

[9] J. L. Gauvan, et al., "Bayesian learning for hidden Markov models with gaussian ,mixture state observation densities," *Speech Communication, Vol.II* pp.205-214, 1992.

[10] R. A. Sukkar, et al., "A two pass classifier for utterance rejection in keyword spotting," *Proceedings of ICASSP 93*, pp.451-454, 1993.

[11] S. Song, "Continuous HMM for word spotting and rejection of non vocabulary word in speech recognition over telephone networks," *Proceedings of EUROSPEECH 93*, pp.1563-1566, 1993.

[12] P. Jeanrenaud, "Phonetic-based word spotter: various configurations and application to event spotting," *Proceedings of EUROSPEECH 93*, pp.1057-1060, 1993.

[13] L. Nguyen, "Search algorithms for software-only real-time recognition with very large vocabularies," *Proceedings of DARPA Human Language Technology Workshop*, pp.91-95, 1993.

[14] S. Austin, et al., "The forward-backward search algorithm," *Prodeedings of ICASSP 91*, pp.697-700, 1991.

[15] 구명완, "N개의 최적문장을 찾을 수 있는 한국어 연속음성인식 시스템," 제11회 음성통신 및 신호처리 워크샵 논문집, pp.48-51, 1994. 10월.

[16] H. Ney, et al., "Improvement in beam search for 10,000-word continuous speech recognition," *Proc. of ICASSP 92*, pp.9-12, 1992.

[17] R. Schwartz and S. Austin, "Efficient, high-performance algorithms for N-best search," *Proc. of DARPA Speech and Natural Language Workshop*, pp.6-11, 1990.

[18] L. R. Rabiner, "The role of voice processing in telecommunications," Proc. of IVTTA 94, pp.1-8, 1994.

[19] M. W. Koo, et al., "An experimental field trial of a large vocabulary speaker independent recognition system," Proc. of IVTTA 94, pp.33-36, 1994.

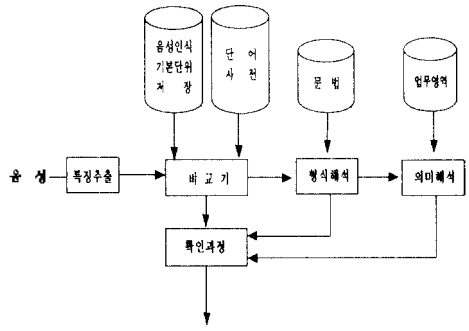

그림 1. 음성인식 시스템의 구성도

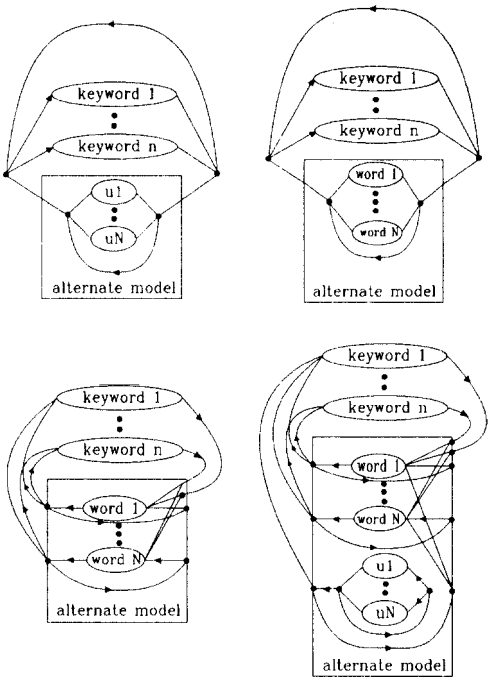

그림2. 여러종류의 FILLER MODEL 방법