자동통역용 한국어 음성 데이타베이스

최 인정 °, 권 오욱, 박 종렬, 김 도영, 정 호영, 은 종관
한국과학기술원 전기및전자공학과

A Korean Speech Database for Use in Automatic Translation

In-Jeong Choi °, Oh-Wook Kwon, Jong-Ryeol Park,
Do-Yeong Kim, Ho-Young Jeong, Chong-Kwan Un
Dept. of Electrical Eng., KAIST

요 약

음성 인식 시스템의 개발을 위해서는 음성 데이타베이스 구축이 중요한 과제의 하나로서, 많은 시간과 노력이 요구된다. 개별적인 음성데이타베이스 구축에 따른 중복 투자를 줄이고 다양한 인식 알고리듬의 성능 비교와 국내 음성 인식 기술의 발전을 위해서는 벤치마크 시험을 위한 공통의 음성 데이타베이스가 필수적이다.

본 논문에서는 한국과학기술원(KAIST) 통신연구실에서 제작한 한국어 음성 데이타베이스에 관하여 기술한다. KAIST 음성 데이타베이스는 자동통역을 위한 무역 상담과 관련된 3,000 단어 규모의 연속어를 비롯하여, 가변 길이 연결 숫자음, phoneme-balanced 75 고립단어, 지역명 관련 500 고립단어, 한국어 아-세트로 구성되어 있다. 이 음성 데이타베이스의 구축을 위하여 사용된 태스크 선정 절차, 녹음 방법, 규격, 및 기대효과 등 세부사항을 기술한다.

I. 서 론

음성 데이타베이스는 음성 인식 및 합성 알고리듬의 연구개발과 알고리듬의 성능 평가에 필수적이다[1]. 선진국의 경우 음성 관련 기술의 발전에 필수적인 공통 음성 데이타베이스 구축의 중요성을 일찌기 인식하여 국가에서 주도적으로 데이타베이스의 구축을 지속적으로 추진하고 있다. 최근 국내에서도 음성 기술 분야의 연구가 본격적으로 시작되고 있어 개발된 여러 인식 알고리듬들의 성능을 객관적으로 평가할 필요가 있다. 그러나 국내의 경우에는 각 연구기관별로 필요시 관련 데이타베이스를 만들어 사용하고 있는 실정이다. ETRI에서 처음으로 4연 숫자음과 고립단어를 공통으로 사용하도록 배포하였으나[2] 현재 널리 활용되지 못하고 있다. 또한 구축된 대부분의 음성 데이타베이스는 고립단어 및 연결 숫자음이 주종을 이루고 있어 대용량 화자독립 연속음성 인식시스템이나 자동통역 시스템을 위한 연속어 공통 데이타베이스는 전무한 실정이다. 결국 공통 음성 데이타베이스의 구축은 연구 개발과 성능 평가의 기준을 마련하는 측면에서의 효과뿐만 아니라 음성기술 관련 연구기관들이 개별적인 음성 데이타베이스 구축을 위해 투자해야 하는 시간과 경비를 크게 절감하는 효과를 함께 가져올 수 있다.

본 연구에서는 공통 음성 데이타베이스의 필요성을 인식하여 체신부의 후원으로 국내 대학 및 연구기관에서 공통으로 사용할 수 있도록 음성 데이타베이스를 구축하였다. 개발된 음성 데이타베이스는 고립단어, 가변길이 연속 숫자음, 연속어 데이타베이스를 포함하고 있다. 특히 연속어 데이타베이스는 대용량 화자독립 연속음성 인식시스템이나 자동통역 시스템의 용도에 부합하도록 설계하였다. 본 논문에서는 개발된 연속어 데이타베이스를 중심으로 데이타베이스의 규격과 구축 절차 및 방법, 그리고 앞으로의 연구 방향에 대하여 기술한다.

II. 음성 데이타베이스의 요건과 연구 경향

이상적인 음성 데이타베이스는 어휘 수나 화자, 녹음환경, 발음형태 등에서 갖추어야 할 요건들이 있다. 특히 대용량 화자독립 음성인식 시스템이나 자동통역 시스템을 위한 음성 데이타베이스는 더욱 엄격한 조건들을 요구한다. 지금까지 선진국에서 개발된 음성 데이타베이스들이 만족시키지 못하고 있지만, 앞으로 지향하고 있는 요구조건들은 다음과 같다. 먼저 어휘의 수에 있어서 대용량이어야 한다. 화자는 연령, 남녀, 지역, 방언, 교육수준 등에 있어서 균형있게 분포되도록 선정되어야 하며, 전화와 같은 실제 환경 조건에서 녹음이 이루어지는것이 합리적이다. 또한 낭독체 음성(read speech)보다는 자연스러운 대화체 음성이어야 한다.

최근 미국, 일본 및 유럽 등의 음성 데이타베이스 구축 사례들을 살펴보면 다음과 같은 측면으로 연구방향이 변화하고 있

음을 알 수 있다.

첫째 어휘의 수와 태스크에서의 변화이다. 80년대에 구축된 음성 데이타베이스들은 한 단위의 어휘 크기와 특정 태스크에 집중되어 있었던 반면, 최근에 구축된 것들은 어휘의 수가 수만에서 수십만에 이르는 텍스트를 이용하며, 특정한 태스크에만 집중되지 않는다는 것이다. 가장 대표적인 사례를 든다면 미국의 WSJ(Wall Street Journal) 음성 데이타베이스[3]와 Le Monde 신문을 텍스트로 사용한 프랑스의 BREF 데이타베이스[4]를 들 수 있다.

둘째 상용시스템 개발을 위하여 많이 이용될 전화 음성 데이타베이스의 개발이다. 현재 6개국의 전화음성 데이타베이스 구축을 위해 진행중인 POLYPHONE 프로젝트가 이러한 변화를 반영하고 있다[5].

세째 실제 환경조건에서의 음성 데이타베이스 구축이다. 이것의 대표적인 사례는 미국의 SRI가 제트비행기 조종실에서 녹음한 ATIS 음성 데이타베이스[6]와 여러 잡음 환경하에서 녹음한 일본의 JEIDA 잡음 음성 데이타베이스[7]가 있다.

네째 다국어 음성 데이타베이스 구축이다. 여러 언어간의 음성기술 개발과 성능 비교의 목적을 위해 개발하고 있으며, POLYPHONE 프로젝트와 미국 OGI 다국어 전화음성 데이타베이스[8]가 좋은 사례가 되고 있다.

III. 무역상담 태스크를 위한 한국어 음성 데이타베이스

현재까지 구축된 음성 데이타베이스는 화자 선정이나 녹음 환경 등에서 여러 제한사항들을 가지고 있으나, 앞으로 계속 보완, 확장해 나갈 예정이다. 음성 데이타베이스는 고립단어, 연결 숫자음, 연속음성 데이타베이스를 포함하고 있으며, 연속음성 데이타베이스를 중심으로 태스크의 선정, 녹음 절차, 및 규격에 대해 설명한다.

3.1 태스크 및 문장 선정

태스크를 선정할 때 고려한 사항은 먼저 자동통역의 용도에 적합하고 어휘가 풍부하며, 대화 형식에 적당해야 한다는 점이다. 이러한 사항들을 고려하여 무역상담을 태스크로 선정하였다.

텍스트를 구성하기 위한 사용된 절차는 다음과 같다. 먼저 무역상담에 관련된 회화책[9]을 참고하여 초기 문장 집합을 구성하였다. 초기 문장 집합은 2,756개의 단어를 사용한 2,210개의 문장으로 구성되었다. 여기서 추출된 단어들을 수작업을 통해 2,293개의 범주(category)로 나누었다. 나누어진 범주를 사용하여 2,150개의 독특한 문장 패턴을 얻을 수 있었다. 그리고 이 과정에서 필요한 단어들을 추가하였다. 추가된 단어들은 시간, 날짜, 숫자, 무역, 지역명 등에 관련된 것과 경어, 존칭어, 시제, 그리고 의미상의 완성을 위한 유사어와 반대어 등으로 이루어져 있다. 추출해낸 독특한 문장 패턴과 category 정보를 이용하여 문장을 랜덤하게 발생시켰다. 문장 발생시의 perplexity는 11.0이었으며, 가능하면 단어들이 고루 나타나고 음소, 음소쌍의 빈도수를 고려하여 중복되지 않는 30,000 문장을 만들었다. 이렇게 만들어진 문장들에서 의미적으로 부적합하거나 발음하기 어려운 문장들과 너무 긴 문장들을 제거하였다. 녹음을 위해 얻어진 문장 집합은 14,300 여개의 문장으로 구성되어 있으며, 3,008개의 단어가 존재하였다. 화자별로 문장을 배분한 방법은 한 화자에 같은 문장 패턴이 포함되지 않도록 하여 100문장씩 분배하였다. 또한 각 화자별로 음소와 음소쌍이 고루 분포되도록 고려하였다.

3.2 화자 및 녹음 환경

녹음한 화자의 수는 총 150명으로서, 남자가 100명, 여자가 50명이었다. 연령 분포를 보면 거의 20대와 30대로 구성되어 있으며, 교육수준은 대부분 대학 재학중이거나 대졸의 학력을 가지고 있다. 화자의 지역별 분포 상황은 서울이나 대전 지역에 많이 분포되어 있으며, 상세한 지역별 분포 상황은 표 1에 나타나 있다.

녹음환경은 조용한 사무실 환경이며, 발음한 음성 신호는 Ariel ProPort 656을 사용하여 16 kHz, 16 bit 선형 PCM으로 A/D 변환되었다. 마이크로폰은 Sennheiser HMD224X headset을 사용하였으며, 컴퓨터는 SPARC10 호환 워크스테이션을 사용하였다.

표 1. 지역별 화자 분포 상황

성별 / 지역	남 자	여 자	계
서울,경기	50	9	59
충청	13	35	48
경상	23	1	24
전라	11	5	16
기타	3	0	3
계	100	50	150

3.3 녹음 방법 및 절차

녹음 과정은 크게 화자에 대한 사전교육, 녹음, 확인 과정, 그리고 데이타베이스에 대한 기록 과정으로 나누어진다.

먼저 녹음을 하기 전에 각 화자에게 음성 데이타베이스 구축의 필요성을 알리며, 녹음 내용을 미리 보여주고 자연스럽게 발음하도록 요청하는 사전교육을 한다. 녹음은 한명의 녹음 관리자가 화자 옆에서 녹음과정을 관리하며 진행된다. 모든 과정은 X 윈도우 환경에서 편리하게 녹음할 수 있도록 되어 있으며, 한 문장씩 읽어 음성 구간만을 검출하는 끝점 검출과정을 거쳐 문장 단위로 한 화일에 저장된다. 화자가 잘못 읽거나 음성 구간의 검출이 실패하면 다시 읽도록 요청하였다. 화일명은 화자

이름과 음성 데이타베이스 종류에 따라 결정되며, 화일명이 중복되지 않도록 주의하였다. 음성 데이타는 일시적으로 시스템의 디스크에 저장하였다가 DAT(Digital Audio Tape Recorder)로 옮겨 저장하였다.

녹음을 모두 마친후 저장된 음성 데이타들을 다시 듣고 수정하는 확인 과정을 거쳤다. 이 과정에서 모든 녹음 문장들을 다시 들어보고 텍스트와 실제 발음이 틀리거나 잡음이 들어있거나, 또는 음성 구간의 검출이 잘못되어 있으면 그 음성 화일을 삭제하였다. 이러한 확인 과정을 거쳐 최종적으로 150명의 화자에 대해 14,746 문장의 발음으로 구성된 음성 데이타베이스를 구축하게 되었다. 화자당 평균 98.3개의 문장을 발음하였으며, 한 문장당 평균 단어수는 8.4단어이다. 전체적으로 문맥의 변이성이 각 화자별로 얼마나 잘 분포되었는가를 나타내는 척도로서 음소쌍이나 트라이폰(triphone) coverage를 사용한다[10]. 이 척도는 각 화자의 발음 문장에서 나타난 음성단위의 수를 전체 화자의 발음 문장에서 나타난 음성 단위로 나눈 값으로 계산된다. 구축된 데이타베이스에서 음소쌍과 트라이폰의 coverage의 평균는 각각 0.55와 0.243, 표준편차는 0.016과 0.010으로 나타났으며, 이러한 분석치로부터 전체적으로 고르게 문맥의 변이성이 분포되었음을 보여주었다. 표 2는 연속어 음성 데이타베이스에서 추출된 규격들을 보여주고 있다.

3.4 음성 데이타베이스의 규격

국내의 음성인식과 자동통역의 연구개발과 성능 평가를 위해 배포할 음성 데이타베이스는 무역상담 연속어, 연결 숫자음, 75 격리단어, 아-세트 격리단어, 500 격리단어 등으로 구성되어 있으며, 각각의 규격은 표 3에 설명되어 있다.

표 2. 연속어 음성 데이타베이스의 규격

항 목	내 용
총 문장수	14,746 문장
사용된 단어수	2986 단어
한 문장당 평균 단어수	8.4 단어/문장
사용된 음성단위의 수	음소 : 40 개 음소쌍 : 909 개 트라이폰 : 6,651 개
평균 음성단위의 coverage	음소 : 0.95 음소쌍 : 0.55 트라이폰 : 0.243

Ⅳ. 결 론

본 논문에서는 국내의 여러 대학 및 연구기관에서 공통으로 사용할 수 있도록 대용량 화자독립 연속음성 인식시스템이나 자동통역 시스템의 응용에 부합하는 음성 데이타베이스를 구축하였다. 개발된 음성 데이타베이스는 무역상담관련 연속어, 가변길이 연결 숫자음, 75 phoneme-balanced 격리단어, 한국어 아-세트, 한국지명관련 500 격리단어 데이타베이스 등 총 5가지로 구성되어 있다. 남자 100명, 여자 50명, 총 150명의 음성을 조용한 사무실 환경에서 녹음하였으며, 녹음된 내용은 회화체 형식의 낭독체 음성이다.

구축된 음성 데이타베이스는 국내의 음성관련 기술의 개발과 성능 평가를 위해 대학이나 관련 연구기관에서 활용할 수 있도록 배포할 예정이다. 현재 구축된 음성 데이타베이스는 앞으로 계속 보완, 확장해 나갈 예정이며, 이 데이타베이스 개발을

표 3. 구축된 음성 데이타베이스의 규격

항 목 / 종 류	내 용	데이터량	녹음 환경
무역상담 연속어 DB	무역상담 연속어 어휘 : 3,000 단어 150명(남100, 여50)	총 14,746 문장 평균 8.4 단어/문장 약 1.5 GByte	조용한 사무실 환경 Ariel ProPort 656 16 kHz, 16 bit Sennheiser HMD224X headset
연결숫자음 DB	3-7자리 연결숫자 어휘 : 11 단어 140명(남90, 여50)	총 5,169 문장 평균 5.1 단어/문장 220 MByte	조용한 사무실 환경 Ariel ProPort 656 16 kHz, 16 bit Sennheiser HMD224X headset
75 격리단어 음성 DB	phoneme-balanced 단어 어휘 : 75 단어 140명(남90, 여50)	총 10,459 단어 207 MByte	조용한 사무실 환경 Ariel ProPort 656 16 kHz, 16 bit Sennheiser HMD224X headset
아-세트 음성 DB	한국어 아-세트 어휘 : 19 단어 140명(남90, 여50)	총 2,647 단어 42 MByte	조용한 사무실 환경 Ariel ProPort 656 16 kHz, 16 bit Sennheiser HMD224X headset
500 격리단어 음성 DB	한국 지명 단어 어휘 : 500 단어 48명(남34, 여14)	총 7,559 단어 178 MByte	방음실 Ariel ProPort 656 16 kHz, 16 bit Hand-hold type

계기로 국내에서 공통으로 사용할 수 있는 음성 데이타베이스 구축을 지속적으로 추진해야 하겠다.

참고 문헌

[1] W. Fisher, V. Zue, J. Bernstein and D. Pallett, "An Acoustic-Phonetic Database," J. Acoustic. Soci. Am., Vol. 81, Suppl. 1, 1987.

[2] 이 용주, 임 연자, 한 남용, 최 준혁, 정 유현, "ETRI의 음성 및 텍스트 데이타베이스의 구축 현황," 제1회 ETRI 음성 언어 및 음향정보처리 워크샵 논문집, pp. 161-177, 1993.

[3] D. B. Paul, J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," Proc. ICSLP 92, pp. 899-902, 1992.

[4] L. F. Lamel, J. Gauvain, M. Eskenazi, "BREF, a Large Vocabulary Spoken Corpus for French," Proc. Eurospeech-93, pp. 505-508, 1993.

[5] J. Bernstein, K. Taussig, "MACROPHONE: An American English Telephone Speech Corpus for the POLYPHONE Project," Proc. IEEE ICASSP-94, pp. I.81-I.84, 1994.

[6] V. Zue, et al., "The MIT ATIS System: Preliminary Development, Spontaneous Speech Data Collection, and Performance Evaluation," Proc. Eurospeech-93, pp. 537-540, 1993.

[7] S. Itahashi, "Recent Speech Database Projects in Japan," Proc. ICSLP 90, pp. 1081-1084, 1990.

[8] Y. K. Muthusamy, R. A. Cole and B. T. Oshika, "The OGI Multi-Language Telephone Speech Corpus," Proc. ICSLP 92, pp. 895-898, 1992.

[9] 이 찬승, 철차별 무역상담 영어, 능률영어사, 1989.

[10] P. Price, W. M. Fisher, J. Bernstein and D. S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," Proc. IEEE ICASSP-88, pp. 651-654, 1988

[11] 은 종관 외, 연속음성 인식시스템 개발 연구 2차년도 최종 보고서, 한국과학기술원, 1994.