KAIST 통신연구실의 음성 데이타베이스 구축 현황

최 인정, 박 종렬, 권 오욱, 김 도영, 정 호영, 운 경관
한국과학기술원 전기 및 전자공학과

On the Present Construction Status of Speech Databases at KAIST Communications Research Laboratory

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

요약

본 논문에서는 한국과학기술원(KAIST) 통신연구실에서 진행중인 한국어 음성 데이타베이스의 개발 현황에 관하여 기술한다. 음성 데이타베이스의 구축을 위하여 사용된 절차와 환경, 및 데이타베이스의 음성학적, 언어학적 성질들이 상세히 기술된다. 데이타베이스는 음성인식 알고리듬의 개발 및 평가를 위하여 사용되도록 고안되었다. 데이타베이스는 5종류의 음성 데이타, 즉 3천단어 규모의 무역관련 연속음성, 가변길이 연결 숫자음, phoneme-balanced 75 고립단어, 지역명 관련 500 고립단어, 한국어 아-세트로 구성되어 있다.

I. 서 론

음성신호는 발성자에 따라 개인차가 심하고 전후에 발성되는 음소의 영향에 의한 조음결합에 따라 그 특성이 크게 변화한다. 이러한 음성의 개인차 및 조음결합의 현상을 분석하기 위해서는 많은 사람이 발성한 다양한 음성 데이타가 필요하다. 또, 각종 음성정보처리 시스템을 개발하기 위해서는 분석, 인식 및 합성의 각종 수단을 적절히 비교 평가해야 하며, 이를 위해서는 공통 음성 데이타를 이용하는 방법 이외에는 없다.

선진국의 경우 음성 관련 기술의 발전에 필수적인 공통 음성 데이타베이스 구축의 중요성을 일찌기 인식하여 국가에서 주도적으로 데이타베이스의 구축을 지속적으로 추진하고 있다. 최근 국내에서도 음성기술 분야의 연구가 본격적으로 시작되고 있어 개발된 여러 알고리듬들의 성능을 객관적으로 평가할 필요가 있다. 그러나 국내의 경우에는 각 연구기관별로 필요시 관련 데이타베이스를 만들어 사용하고 있는 실정이다. 또한 구축된 대부분의 음성 데이타베이스는 고립단어 및 연결 숫자음이 주종을 이루고 있어 대용량 화자독립 연속음성인식 시스템이나 음성대화 시스템을 위한 연속어 공통 데이타베이스는 전무한 실정이다. 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 논문에서는 한국과학기술원 통신연구실에서 진행중인 한국어 음성 데이타베이스 구축현황에 관하여 기술한다. 현재까지 구축된 음성 데이타베이스는 고립단어, 가변길이의 연결 숫자음, 3천단어 규모의 연속음성 데이타베이스를 포함하고 있으며, 연속음성 데이타베이스를 중심으로 태스크의 선정, 구축 절차와 환경, 및 규격에 대해 설명한다.

II. 음성 데이타베이스의 구축 방법 및 규격

A. 태스크 및 문장 선정

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

텍스트는 초기 문장집합 구성, 단어 조정, 문장의 발생, 화자별 문장 배분 등의 절차를 거쳐 구성되었다. 초기 문장집합은 무역상담에 관련된 회화책[1]을 참고하여 구성하였다. 이렇게 얻어진 문장 집합에서 무역상담과 무관한 문장들은 삭제하였다. 너무 구어체적인 문장들은 삭제하거나 문어체 형식으로 변환하였다. 또한 너무 긴 문장들은 삭제하거나 더 짧은 문장들로 분리되었다. 얻어진 초기 문장집합은 2,756개의 단어를 사용한 2,210개의 문장으로 구성되었다.

단어 조정 과정에서는 추출된 단어들을 단어 클래스로 분리하고, 태스크의 완성을 위해 빠진 단어들을 추가하였다. 여기서 단어 클래스는 단어의 형태소적 범주와 의미상 범주를 고려하여 단어들을 분류한 것이다. 수작업에 의해 추출된 2,756개의 단어들은 2,203개의 단어 클래스로 분리 되었다. 시간, 날짜, 숫자, 지역명 등과 관련된 단어들과 무역에 관련된 단어들이 주로 추가되었다. 또한 경어와 존칭어, 시제, 그리고 의미상의 완성을 위한 유사어와 반대어 등의 단어들을 추가하였다. 결국 문장의 생성을 위해 3,008개의 단어가 사용되었다.

문장은 3,008개의 단어와 단어 클래스 정보를 이용하여 발생되었다. 단어열로 되어 있는 초기 문장집합을 단어 클래스 열로 바꾼 후, 중

복되는 문장 패턴을 제거하여 2,150개의 독특한 문장 패턴을 추출하였다. 여기서 얻어진 독특한 문장 패턴과 단어 클래스 정보를 이용하여 랜덤하게 문장을 발생시켰다. 문장 발생시의 perplexity는 11.07이었다. 여기서 perplexity란 정성적으로 임의의 단어뒤에 올 수 있는 평균적인 후보 단어의 수라고 할 수 있다[10]. 가능하면 단어들이 고루 나타나고 음소, 음소쌍의 빈도수를 고려하여 중복되지 않도록 30,000 문장을 만들었다. 이렇게 만들어진 문장중에서 의미적으로 부적합하거나 발음하기 어려운 문장들과 너무 긴 문장들은 제거하였다. 녹음을 위해 얻어진 문장 집합은 14,300여개의 문장으로 구성되어 있으며, 3,008개의 단어가 존재하였다.

마지막으로 화자별 문장 배분 과정에서는 한 화자에 같은 문장 패턴이 포함되지 않도록 하여 100문장씩 배분하였다. 또한 각 화자별로 음소와 음소쌍이 고루 분포되도록 고려하였다.

B. 화자 및 녹음 환경

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

표 1. 지역별 화자 분포 상황(단위 : 명)

지역 성별	남자	여자	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

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.

C. 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. 녹음을 위해 한명의 관리자가 화자 옆에서 녹음과정을 관리하며, 모든 과정은 X 윈도우 환경에서 편리하게 녹음할 수 있도록 되어 있다. 음성은 한 문장씩 읽어 음성 구간만을 검출하는 끝점검출 과정을 거쳐 문장 단위로 한 화일에 저장된다. 화자가 잘못 읽거나 음성 구간의 검출이 실패하면 다시 읽도록 요청하였다. 화일명은 화자 이름과 남녀의 구분, 음성 데이타베이스 종류에 따라 결정되며, 화일명이 중복되지 않도록 주의하였다. 녹음된 음성 데이타는 일시적으로 시스템의 하드 디스크에 저장하였다가 DAT(Digital Audio Tape)로 옮겨 저장하였다.

녹음을 모두 마친 후 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.

마지막으로 구축된 음성 데이타베이스에 관련된 사항들을 기록하였다. 이 과정에서 데이타베이스의 녹음 환경, 녹음 방법 및 화자 정보 등의 정보를 편집하고, 전체적인 데이타베이스의 구조를 완성하였다.

D. 음성 데이타베이스 규격

구축한 음성 데이타베이스는 150명의 화자가 발음한 14,746 문장으로 구성되어 있으며, 이것은 총 12시간 24분(44,640초) 동안의 음성 데이타량에 해당된다. 화자당 평균 98.3개의 문장을 발음하였으며, 한 문장당 평균 단어수는 8.4 단어이다. 발음속도는 평균 일분당 166.5 단어이었다.

전체적으로 문맥의 변이성이 각 화자별로 얼마나 잘 분포되었는가를 나타내는 척도로서 음소쌍이나 트라이폰(triphone) coverage를 사용한다. 이 척도는 각 화자의 발음 문장에서 나타난 음성단위의 수를 전체 화자의 발음 문장에서 나타난 음성단위로 나눈 값으로 계산된다. 구축된 데이타베이스에서 음소쌍과 트라이폰의 coverage의 평균은 각각 0.55와 0.243, 표준편차는 0.016과 0.010으로 나타났으며, 이러한 분석치로부터 전체적으로 고르게 문맥의 변이성이 분포되었음을 알 수 있다. 표 2는 연속어 음성 데이타베이스에서 추출된 규격들을 보여주고 있다.

표 2. 연속어 음성 데이타베이스의 규격

항 목	내 용
총 문장수	14,746 문장
사용된 단어수	2,986 단어
한 문장당 평균 단어수	8.4 단어/문장
총 발성시간(duration)	44,640 초
평균 분당 단어수(발음속도)	166.5 단어/분
한 문장당 평균 발음 시간	3.03 초/문장
사용된 음성 단위의 수	음 소 : 40 개 음 소 쌍 : 909 개 트라이폰 : 6,651 개
평균 음성단위의 coverage	음 소 : 0.95 음 소 쌍 : 0.55 트라이폰 : 0.24

「표준 한국어 발음 대사전」(KBS저, 어문각, 1993)의 통계에 따르면 한국어에서 한 낱말의 평균 음절의 수는 약 2.87개이며, 평균 음소의 수는 7.55개이다. 또, 한 음절당 음소의 수는 2.63개이다. 한국어의 음절 형태는 모음을 V, 자음을 C라 할 때, V 형, VC 형, CV

형, CVC 형 등 네가지 형태가 존재한다. 이론적으로 한국어에서는 모두 3,520가지의 음절 종류가 가능하나, 실제로 쓰이는 가짓수는 음소의 연결에 제약이 있기 때문에 이보다 훨씬 적다. 그러나 아직 아무도 우리말에 실제로 쓰이는 음절의 가짓수가 몇 개인지 정확하게 밝혀 내지는 못했다. 참고로 「표준 한국어 발음 대사전」에 쓰인 음절의 가짓수는 1,153개로 조사되었다. 구축된 음성 데이타베이스에 대해 조사한 결과, 사용된 음절의 가짓수는 662개이었다. 한 단어의 평균 음절의 수는 약 2.89개, 평균 음소의 수는 9.17개, 그리고 한 음절당 음소의 수는 2.78개로 나타났다.

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.

E. 문법적 특성

텍스트를 구성할 때 사용된 단어의 수는 3,008개, 단어 클래스의 수는 2,293개였으며, 문장 생성시의 perplexity는 11.07이었다. 그러나 연속음성인식 시스템이나 자동통역 시스템의 개발을 위해서는 더 완화된 문법적 제약이 필요하다. 더우기 문어적인 표현과는 달리 구어적인 표현들은 정형화된 문법적 룰을 벗어나는 경향이 심하다.

본 논문에서는 단어 클래스에 근거한 문법을 작성하였다. 이 모델은 먼저 각 단어들을 단어 클래스에 할당한 후 단어 클래스에 근거한 가능한 문맥을 찾아 확률값을 부여한다. 단어 클래스에 의한 언어 모델은 문법적 제약을 위해 필요한 파라미터의 수가 적으므로, 학습 데이타를 효율적으로 활용할 수 있다는 장점이 있다. 단어 클래스는 원래 명사나 형용사, 관형사 등과 같은 형태소적 범주와 회사명, 국가명 등과 같은 특수한 의미상의 범주 등에 의해 분류된다. 한국어에서는 문법의 룰이 불완전하고, 특히 구어적 표현에서는 더욱 심하므로 본 논문에서는 회사명, 선적명 등 무역상담 태스크에서의 특수한 의미상 범주와 날짜, 숫자 등과 같은 일반적 단어 범주 등을 주로 사용하였다. 또한 단어 범주를 세분화하기 위하여 주어진 단어와 앞뒤에 올 수 있는 단어 클래스들을 조사하여, 비슷한 상황에 있는 단어들을 하나의 단어 범주로 묶어준다. 초기에 단어 범주들이 텍스트에 심하게 의존하지 않도록 수작업을 통해 비슷한 성격의 단어들을 클래스로 모아 주었다. 초기의 단어 클래스를 시작으로 하여 주어진 단어 클래스에 인접할 수 있는 단어 클래스의 종류를 조사하고, 다른 단어 클래스의 상황과 비교하여 그 정합 정도를 계산하며, 주어진 단어 클래스의 발생 빈도수를 고려하여 최종적인 정합의 정도를 정량화한다. 정합의 정도는 두 단어 클래스의 앞뒤에 올 수 있는 클래스의 전체수에 대한 같은 위치에서 두 단어 클래스 모두에 나타난 클래스 수의 비로 측정된다. 발생 빈도수가 적은 단어 클래스들을 우선적으로 묶어주기 위해 발생 빈도수의 역수를 가중치로 곱한다. 정합의 정도가 가장 큰 단어 클래스 쌍을 찾아 같은 단어 클래스로 묶어준다. 이러한 과정을 반복하여 원하는 단어 클래스 수나 다른 수렴 조건이 만족될 때까지 계속한다. 표 4는 단어 클래스의 수와 word pair 문법 적용시의 perplexity 값의 변화를 보여주고 있다.

표 4. 단어 클래스 수와 Perplexity

단어 클래스의 수	Perplexity (word pair 문법)
500	644.0
800	143.5
1,000	51.4
1,200	28.2
1,500	22.7
2,920	6.3

표 3. 구축된 음성 데이타베이스의 규격

종 류	내 용	데이타량	녹음 환경
무역상담 연속어 DB	무역상담 연속어 어휘 : 약 3,000 단어 150명(남100,여50)	총 14,176 문장 평균 8.4 단어/문장 약 1.5 GBytes	조용한 사무실 환경 Ariel ProPort 656 16 kHz, 16 bit Sennheiser HMD224X headset
연결숫자음 DB	3-7 자리 연결숫자 어휘 : 11 단어 140명(남90,여50)	총 5,169 문장 평균 5.1 단어/문장 220 MBytes	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 MBytes	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 12 MBytes	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 MBytes	방음실 Ariel ProPort 656 16 kHz, 16 bit Hand-hold type

III. 결 론

지금까지 음성정보처리 연구에 있어서 가본적인 연구개발 도구인 동시에 개발 내용의 객관적인 평가의 기준이 되는 음성 데이타베이스에 관하여 기술하였고, 한국과학기술원 통신연구실에서 구축한 한국어 음성 데이타베이스에 대하여 논하였다. 개발된 음성 데이타베이스는 무역상담관련 연속어, 가변길이 연결 숫자음, 75 phoneme-balanced 격리단어, 한국어 아-세트, 한국지명관련 500 격리단어 데이타베이스 등 총 5가지로 구성되어 있다. 개발된 음성 데이타베이스들은 충분한 협의 과정을 거쳐 국내 음성관련 연구기관과 대학이 자유롭게 이용할 수 있도록 배포할 예정이다. 구축한 음성 데이타베이스는 화자 선정, 발성내용, 녹음 조건 등에서 제한사항들을 가지고 있지만, 공통으로 이용 가능한 음성 데이타베이스가 없는 국내 상황에서 각종 음성기술의 개발내용에 대한 객관적인 평가기준을 제공하는데 일조할 것으로 생각된다.

자동통역 시스템의 개발을 위해서 아직 상당한 제한 요소들을 지니고 있는 현재의 음성 데이타베이스는 다음과 같은 방향으로 보완, 확장되도록 연구가 진행되어야 할 것이다. 첫째, 발성자의 영역을 확대하고, 녹음 조건을 더 실제화해야 한다. 둘째, 조음결합과 같은 영향들을 잘 표기할 수 있도록 음운적 균형이 이루어진 텍스트와 구성이 이루어져야 한다. 이를 위해서는 대규모의 통계적 조사와 알고리듬들의 개발이 뒤따라야 한다. 셋째, 자동으로 레이블링(labelling)할 수 있는 기법의 개발이다. 음성 신호의 연구를 위해서는 레이블링이 필수적이나 수동으로 레이블링 할 경우 많은 시간이 필요하며, 또한 작업을 수행하는 사람의 능력에 따라 다른 결과가 나타나게 된다. 이러한 문제들을 줄이기 위해 축적된 know-how와 알고리듬 개발을 통해 자동적으로 레이블링할 수 있는 기법의 개발이 필요하다. 넷째, 문법과 같은 언어처리와 언어지식에 관한 연구가 이루어져야 한다. 이를 위해서는 대규모의 텍스트 데이타베이스의 개발이 요구된다. 마지막으로 융이하게 이하는 상황의 음성 데이타를 찾을 수 있는 데이타베이스의 구조에 관한 연구가 이루어져야 할 것이다.

음성 데이타베이스는 구축시 많은 시간과 노력을 필요로 하기 때문에 국내에서는 아직 공동으로 이용 가능한 음성 데이타베이스에 관한 연구가 본격적으로 이루어지지 못했다. 따라서 국내 음성정보처리 연구의 저변확대 및 활성화를 위하고 동시에 음성 기술의 개발 내용에 대한 객관적인 평가 기준을 제공하는 음성 데이타베이스의 구축에 관한 국가 차원의 집중적이고 지속적인 연구개발이 이루어져야 할 것이다.

참고 문헌

1. 이 찬승, 실차별 무역상담 영어, 능률영어사, 1989.

2. M. Phillips, J. Glass, J. Polifroni and V. Zue, "Collection and Analyses of WSJ-CSR Corpus at MIT," *Proc. ICSLP 92*, pp. 907-910, 1992.

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. L. Gauvain, L. F. Lamel, and M. Eskenazi, "Design Considerations and Text Selection for BREF, a Large French Read-Speech Corpus," *Proc. ICSLP90*, pp. 1097-1100, 1990.

6. 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.

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

8. S. Itahashi, "Recent Speech Database Projects in Japan," *Proc. ICSLP 90*, pp. 1081-1084, 1990.

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

10. F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of IEEE*, Vol. 64, No. 4, pp. 532-556, Apr. 1976.