음성정보기술산업지원센터의 음성 코퍼스 구축 현황 및 계획

김봉완*, 이용주**
원광대학교 SITEC*, 원광대학교 전기전자 및 정보공학부**

Current States and Future Plans for Speech Corpora at SITEC

Bong-Wan Kim*, Yong-Ju Lee**
SITEC, Wonkwang Univ.*
Div. of Electric and Electronic Eng., Wonkwang Univ.**
{bwkim, yjlee}@sitec.or.kr

요약

최근 컴퓨터와 인간간의 대화 수단으로 음성을 활용하는 기술인 음성정보기술이 발달함에 따라 대어휘 연속 음성 인식 및 무제한 어휘 음성 합성의 고도화를 위한 연구가 진행되고 있다. 음성 합성의 경우에도 최근 대형의 음성 데이터 베이스로부터 임의 길이의 음성 부분을 골라내어 접속함으로써 좋은 합성 품질을 얻고 있다. 따라서 이러한 연구에 사용될 음성 코퍼스에 관한 요구와 관심이 높아지고 있다. 본 논문에서는 음성정보기술산업지원센터(SITEC)에서 구축중인 음성 코퍼스의 현황과 향후 계획에 관하여 보고한다. 방음실환경에서의 인식 및 합성 연구용 코퍼스, 아동용 음성 코퍼스, Dictation용 음성 코퍼스, 자동차내 소음 및 음성 코퍼스 등의 구축 내용이 소개된다.

1. 서론

한국어의 공학적인 응용을 위해서는 그 기반이 되는 요소기술로써 음성인식 및 합성으로 대표되는 음성 처리기술과 언어 이해 및 기계번역으로 대표되는 언어 처리 기술의 연구가 필요하다. 이러한 음성 및 언어 처리기술의 연구를 위해 가장 먼저 확보되어야 할 것이 음성, 언어 및 각종 사전 코퍼스 등 국어 정보베이스이다. 이들의 체계적인 조기확보 여하에 따라 음성 및 언어처리연구의 성패를 좌우한다고 해도 과언이 아니다. 특히 한국어 음성을 대상으로 한 음성 코퍼스는 음성 언어 연구의 기본으로서 개발초기부터 확보되어야 할 연구자원이다[1].

따라서 음성 및 언어의 연구를 위해 데이터가 중요하다는 것은 연구자간에 이론이 없으나 공동으로 사용하기 위한 대규모의 데이터베이스를 갖추는 일은 간단한 문제가 아니다. 본질적으로 데이터는 연구 그 자체의 일부이며 어떤 데이터를 어떻게 모아 가공하는가는 연구 내용에 크게 의존한다. 따라서 다른 연구 목적간에 두루 쓰일 수 있도록 노력하여야하나 완벽하게 범용일 수는 없다. 그러나 대량의 데이터가 공통의 포맷으로 제공되는 것만으로도 그 의의는 매우 크다.

본 논문에서는 음성정보기술산업지원센터(SITEC)에서 구축중인 음성 코퍼스의 현황과 향후 계획에 관하여 보고한다. 방음실환경에서의 인식 및 합성 연구용 코퍼스, 아동용 음성코퍼스, Dictation용 음성 코퍼스, 자동차내 소음 및 음성 코퍼스 등의 구축 내용이 소개된다.

2. SITEC의 음성코퍼스 구축 현황

본 장에서는 1차년도에 진행된 음성 코퍼스 구축 현황에 대하여 기술한다. 현재 음성 코퍼스는 1차년도 구축 계획에서 계획된 음성 코퍼스가 모

두 데이터 수집이 완료된 상태이며 일부 코퍼스에 대해서는 이에 대한 편집, 검증 및 후처리 작업이 진행 중에 있다.

2.1 산업응용 기반 기술 기초 연구용 코퍼스

산업응용 기반 기술 기초 연구용 코퍼스의 발성 목록은 다양한 응용과 연구에 적용하기 위하여 특정 응용에 종속되지 않은 발성목록을 사용하는 것이 바람직하다. 따라서 센터에서는 이러한 목적으로 사용하기 위해 PRW 4,178어절을 선정하고 이를 발성목록으로 사용하였다.

선정된 PRW 4,178어절은 한국어에서 발생할 수 있는 다양한 음운 환경을 고려하고 있으며 또한 음절에 대한 고려도 함께 이루어진 발성 목록이다. 이렇게 구성된 발성 목록을 10개의 세트로 나누어 발성하도록 하였다.

센터의 지역협력 사이트를 활용하여 전국적으로 500명 화자의 음성 데이터를 방음실에서 수집하였으며 남녀 성비는 1:1이다. 1인당 발성량은 417 ~ 418단어를 발성하였다.

또한 수집된 음성 데이터 전량에 대하여 음운 레이블링 기준(센터 권고안)에 의해 자동 레이블링을 완료하였으며, 자동 레이블링 결과의 검증을 위해 지역 협력 사이트에서 레이블링 전문 인력에 의해 검증 및 수정 작업을 진행중이다.

2.2 아동용 음성 코퍼스

최근 완구, 교육용 S/W 등 아동용 응용에 대한 요구가 증가함에 따라 아동용 음성 인식 응용을 위한 음성 코퍼스를 구축하였다. 다양한 응용의 개발을 위한 발성 목록의 구성은 다음과 같다.

- 4연 숫자 : 340종
  - 앞, 뒤의 다양한 숫자 환경을 포함한 목록
- PBW : 452 종[2]
  - 한국어의 다양한 음운 환경을 포함한 발성 목록
- 명령 및 지시어 : 400종
  - 컴퓨터, 통신 등의 응용을 위한 명령 및 지시어
- 단독 숫자 : 41종
  - 단독 숫자 및 단위

위와 같이 구성된 발성 목록은 1,233종으로 1명이 아동이 발성하기에는 그 양이 너무 많다. 따라서 단독 숫자의 경우 모든 아동이 발성하도록 하고 나머지 4연 숫자, PBW 와 명령 및 지시어는 20개의 세트로 나누어 구성하고 각 아동은 1개의 세트만을 발성하도록 하였다.

총 500명의 초등학교 학생을 대상으로 데이터를 수집하였으며 남녀 성비는 1:1이며, 1인당 발성량은 100 ~ 101단어이다. 수집 환경은 사무실 또는 가정집에서 PC의 사운드카드와 Andrea ANC 750마이크를 이용하여 수집하였다.

2.3 Dictation용 낭독 음성 코퍼스

(가) 발성 목록 설계

발성 목록 선정을 위한 모집단으로는 KAIST에서 구축된 4,300만 어절의 KAIST Corpus를 사용하였다. 그러나 이러한 텍스트 코퍼스는 그 내용이 문자언어이므로 자연스럽게 낭독하는데는 곤란한 표현이 존재할 수 있다. 따라서 이러한 문장은 발성목록 후보에서는 제외하고 고빈도 어휘에 대한 분석을 수행하였다.

분석 결과 상위 고빈도 5,000어절이 전체 어절에 대해 50.6%의 coverage를 가지며 상위 10,000어절의 경우 58.4%, 20,000어절의 경우 66.2%의 coverage를 갖는 것으로 나타났다. 상위 고빈도 77,000어절의 경우 약 80%의 coverage를 갖게되나 이러한 경우 어휘의 수가 너무 많아지게 된다. 따라서, 본 코퍼스에서는 발성 목록 선정을 위해 고빈도 5,000어절, 8,000어절 및 10,000어절을 발성 목록 선정을 위한 대상어휘로 선정하였으며, 향후확장할 예정이다.

추출된 문장의 총 수는 20,833문장으로 문장의 평균 길이는 문장 당 7.43어절이다.

또한 인식 대상 어휘에 포함되지 않은 단어가 발성된 경우에 대처하기 위한 OOV(Out of vocabulary) 테스트를 위해 다음과 같이 문장 목록을 구성하였다.

- 5K 문장 세트 (8,608 문장)
  - 고빈도 5,000어절에 포함된 어휘만으로 구성된 문장 세트
- 8K-5K 문장 세트 (7,301 문장)
  - 고빈도 8,000어절에 포함된 어휘만으로 구성된 문장 세트을 구성하고, 여기에서 5K

문장 세트에 포함된 문장은 중복되므로 이를 삭제한 것

- 10K-8K 문장 세트 (4,924 문장)
  - 고빈도 10,000어절에 포함된 어휘만으로 구성된 문장 셋을 구성하고, 여기에서 5K 문장 세트와 8K-5K 문장 세트에 포함된 문장은 중복이므로 이를 삭제한 것

위와 같이 구성된 총 20,833 문장은 한 사람이 모두 발성하기에는 그 양이 너무 많다. 따라서 위의 문장을 1인당 평균 104.17문장을 발성할 수 있도록 200개의 발성 세트로 재구성하였으며 각각의 세트에는 위에서 기술된 세 가지의 문장 세트가 골고루 분포하도록 구성하여 모든 화자가 세 가지 세트에 포함된 문장중 일정한 양을 발성할 수 있도록 배려하였다.

(나) 음성코퍼스의 특징

남, 녀 각 200명의 화자에 대하여 음성 데이터를 수집하였으며 1인당 발성량은 104 ~ 105문장이다. 데이터는 사무실 환경에서 PC의 사운드카드와 Andrea ANC 750 마이크를 이용하여 수집되었다.

2.4 수출 지원을 위한 외국어 음성 코퍼스 : 중국어 음성 코퍼스

(가) 발성 목록

수출 지원을 위한 외국어 음성 코퍼스로 1차년도에는 중국어 음성 코퍼스를 구축하였다. 발성 목록은 다음과 같이 2,648개의 단어와 400개의 문장으로 구성되어 있다.

- 음절
  - 421개의 음절
  - 중국어 성조를 모델링하기 위해 경성 및 4성을 고려한 음절 수집
  - 수집된 음절의 수 : 421 X 5 = 2,105 음절
- PBW 단어
  - 2음절 ~ 4음절의 1,200단어
- 사연숫자
  - 661개의 사연숫자
- 날짜 관련 단어
  - 366개의 날짜 관련 단어
- 문장
  - 각 문장마다 최대 27개의 한자(음절)를 포함하는 400개 문장

(나) 음성 코퍼스의 특징

연변대학 지역협력 사이트를 통하여 북경어를 사용하는 화자를 중심으로 화자를 모집하였으며, 방언을 사용하는 화자는 가능한 배제하였다. 총 300명의 화자가 발성하였으며 성비는 2:3으로 구성되어 있다. 1인당 발성량은 110 ~ 123단어와 20문장으로 구성되어 있다. 데이터는 사무실 환경에서 PC의 사운드카드를 통하여 Sennheiser E835 마이크를 통하여 수집되었다.

2.5 산업 응용을 위한 운율 합성용 음성 코퍼스

(가) 발성 목록 설계

음성 합성 시스템을 위한 발성 목록 선정의 모집단으로 사용된 텍스트 코퍼스는 설명문, 수필문, 사회학, 방송 3社(KBS, MBC, SBS 등)의 뉴스, 신문(조선일보, 한국일보), 경제학, 전산학, 기계학, 생물학 등의 장르별 균형 텍스트로 구성된 KAIST Taged Corpus 100만 어절을 사용하여 Triphone maximization 기준을 적용하여 발성 목록을 선정하였다.

선정된 문장 발성 목록은 4,360문장이며 이 문장에 포함된 Triphone의 총 종류수는 모집단과 같이 18,025 종류이다.

(나) 음성 코퍼스의 특징

남, 녀 전문 성우 각 1인이 방음실에서 발성하였다. 마이크는 Rode NT-2를 사용하였으며 EGG 신호도 동시에 수집되었다. 수집된 음성데이터 중 여성화자의 1000문장에 대하여 음운 레이블링을 완료 하였으며, K-ToBI (Korean-Tone and Break Index) 기준(Ver 3.1)을 적용하여 운율 레이블링을 실시하고 있다.

2.6 자동차 음성 코퍼스 및 소음 코퍼스 프로토타입

최근 자동차 환경에서의 음성인식 응용에 대한 관심과 수요가 많아지고 있다. 산업자원부에서도 중기거점과제를 통하여 자동차 환경에서의 음성인식기술에 대해 지원하고 있다. 따라서, 센터에서는 이러한 환경에서의 음성 인식 기술 연구 및 개발을 위해 반드시 필요한 자동차 소음코퍼스, 음성 코퍼스를 구축하기로 결정하였다. 그

러나 자동차 환경에서의 소음 및 음성 코퍼스의 경우 그 수집 절차, 환경 요인 등에 있어서 일반적인 경우와 달리 매우 많은 변수가 있다고 할 수 있다. 따라서 1차년도에는 이러한 수집 절차 및 환경 요인에 대한 연구와 분석을 위한 프로토타입 코퍼스를 구축하였다.

구축된 코퍼스는 소음 환경으로 자동차 요인, 도로 요인등을 고려하여 270종의 환경을 정의하고, 각 환경에 대하여 동시에 8개의 채널을 통하여 소음 데이터를 수집하였다. 음성 코퍼스의 경우 80km의 주행상황으로 환경을 한정하고 100명의 화자에 대한 음성을 8개의 채널을 통하여 수집하였다.

2.7 다양한 환경의 시험용 코퍼스

음성 인식 시스템의 성능에 영향을 미치는 다양한 요인 중 마이크의 음향적 특성, 마이크의 위치 및 마이크와 화자와의 거리도 매우 중요한 요인중 하나이다. 따라서 센터에서는 이러한 다양한 변인에 따른 시험용 코퍼스를 구축하기로 하였으며 1차년도에는 마이크의 종류에 따른 시험용 코퍼스를 구축하였다.

수집절차는 방음실에서 고성능 다이크로폰으로 수집한 PBW 452단어 70명분이 2회 발성한 코퍼스를 대상으로 HATS(Head And Torso Simulator)를 이용하여 데이터를 수집하였다. 1차년도에는 마이크의 종류별 특성을 살펴보기 위해 해외 Headset microphone 4종, 스텐드형 마이크 4종 등 총8종을 대상으로 방음실 환경에서 데이터를 수집하였다.

2.8 기존 음성 코퍼스의 공유 유도 현황

기존에 구축된 음성 코퍼스 중 센터를 통하여 공유의사를 표명한 음성코퍼스는 다음과 같다.

- PC 환경 숫자음 500명분
- PRW 전화음성 2000명분
- 숫자음 전화음성 2000명분
- 클린스피치 PBW 70명분
- 클린스피치 PBS 20명분
- KAIST 무역상담 코퍼스 외 5종

3. 향후 계획

센터에서는 향후 현재 구축된 음성코퍼스의 내용과 양을 지속적으로 보완하고 확장할 계획이다. 특히 현재 구축된 자동차 음성 코퍼스 프로토타입을 바탕으로 관련 기관 및 연구자들과의 다각적인 논의를 통하여 스펙을 결정하고 대규모의 자동차 음성 코퍼스를 구축하고자 한다. 또한 수출 지원을 위한 외국어 음성 코퍼스 구축을 위하여 영어와 스페인어 음성 코퍼스의 구축, Embeded system을 위한 음성 코퍼스의 구축, 숫자음 음성 코퍼스의 보완, Dictation 음성 코퍼스의 보완 등을 계획하고 있다.

센터는 이러한 음성 코퍼스의 구축 내용과 방향에 대한 관련 연구자들의 많은 참여와 의견을 기대하고 있으며, 제시된 의견을 적극 반영하고자 한다.

4. 결론

본 논문에서는 SITEC에서 구축중인 음성코퍼스의 현황과 향후 계획에 관하여 보고하였다. 방음실환경에서의 인식 및 합성 연구용 코퍼스, 아동용 음성코퍼스, Dictation용 음성코퍼스, 자동차내 소음 및 음성 코퍼스 등의 구축 내용이 소개되었다. 센터에서는 향후 현재 구축된 음성코퍼스의 내용과 양을 지속적으로 보완하고 확장할 계획이며, 음성 코퍼스의 구축 내용과 방향에 대한 관련 연구자들의 많은 참여와 의견을 기대하고 있다.

5. 참고 문헌

[1] 이용주, "음성언어코퍼스," 한국정보과학회지, 1998.2

[2] 김봉완, 이용주 외, "공동이용을 위한 음성코퍼스의 설계 및 구축에 관한 연구," 한국음향학회지, 16권 4호, 1997

[3] 최기선, KAIST 언어자원 2001년도판, 과학기술부 핵심 소프트웨어 과제 결과물 1995-2000 (http://kibs.kaist.ac.kr)

[4] Yong-Ju Lee, Bong-Wan Kim, Yongnam Um, "Speech Information Technology & Industry Promotion Center : Activities and Directions," Proc. of LREC 2002, pp. 1851-1854, 2002. 5.