SiTEC의 공동 이용을 위한 음성 코퍼스 구축 현황 및 계획

김봉완*, 최대림*, 김영일*, 이광현*, 이용주**
원광대학교 SiTEC*, 원광대학교 전기전자 및 정보공학부**

Current States and Future Plans at SiTEC for Speech Corpora for Common Use

Bong-Wan Kim*, Dae-Lim Choi*, Young-Il Kim*, Kwang-Hyun Lee*, Yong-Ju Lee**
SiTEC, Wonkwang Univ.*
Div. of Electric and Electronic Eng., Wonkwang Univ.**
E-mail : {bwkim, dlchoi, yikim, khlee, yjlee}@sitec.or.kr

Abstract

음성정보기술 산업을 효과적으로 지원하기 위해서는 상품 및 기술의 개발을 위한 표준화된 음성 코퍼스의 구축 및 보급이 필수적이라고 할 수 있다. 본 논문에서는 음성정보기술산업지원센터(SiTEC)의 1~2차년도(2001 . 5. 1 ~ 2003. 4. 30)의 사업기간 중에 구축된 음성 코퍼스의 현황 및 향후 계획을 소개한다. 전통산업분야에 대한 음성정보기술 적용확산을 위한 자동차 소음 및 대규모 다채널 자동차 음성 코퍼스, 수출지원을 위한 다양한 외국어 음성 코퍼스, 방음실 환경에서의 인식 및 운율 합성 연구용 코퍼스 , Dictation용 음성 코퍼스, 아동용 음성 코퍼스 등의 구축 내용이 소개된다.

I. 서론

한국어의 공학적인 응용을 위해서는 그 기반이 되는 요소기술로써 음성인식 및 합성으로 대표되는 음성 처리기술과 언어 이해 및 기계번역으로 대표되는 언어처리 기술의 연구가 필요하다. 이러한 음성 및 언어처리기술의 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]. 기관별로 자체 연구를 목적으로 개별적인 코퍼스 구축이 이루어져 왔으나, 최근에는 공동으로 사용할 수 있는 우리말 음성언어 코퍼스의 중요성을 인식하고 체계적인 구축 및 공개에 대한 노력을 시작하였고 이에 대한 산업자원부의 정책적인 지원에 의해 음성정보기술산업지원센터(SiTEC)가 설립되어 활동 중이다.

II. SiTEC의 음성코퍼스 구축 현황

본 장에서는 산업자원부의 산업기반기술조성사업의 일환으로 음성정보기술산업지원센터에서 1~2차년도 사업기간 중에 구축된 음성 코퍼스에 관한 사항을 상세히 기술하고 3차년도 음성 코퍼스 구축 계획에 관해 소개한다. 구축된 음성 코퍼스는 세밀한 검증을 거쳐 2002년 8월부터 보급을 시작하여 현재 총 16종의 음성 코퍼스를 배포하고 있고, 센터 홈페이지를 통해 자세한 사양과 샘플 데이터를 제공하고 있으며, 조만간 8종의 음성 코퍼스를 추가 보급하기 위해 준비 중이다[2]. 아울러, 사용자가 원하는 사양의 음성 코퍼스를 별도 제작하거나, 보급중인 코퍼스 중 원하는 사양의 데이터만을 골라 재작성하는 맞춤형 음성 코퍼스 보급도 준비하고 있다.

2.1 다채널 자동차 소음 및 음성 코퍼스

2.1.1 자동차 소음 및 음성 코퍼스 프로토타입

최근 자동차 환경에서의 음성인식 응용에 대한 관심과 수요가 많아지고 있고, 산업자원부에서도 중기거점과제를 통하여 자동차 환경에서의 음성인식기술의 개발을 지원하고 있다. 자동차 환경에서의 소음 및 음성 코퍼스의 경우 그 수집 절차, 환경 요인 등에 있어서 일반적인 경우와 달리 매우 많은 변수가 있기 때문에 1차년도(2001. 5. 1 ~ 2002. 4. 30)에는 이러한 수집 절차 및 환경 요인에 대한 연구와 분석을 위한 프로토타입 코퍼스를 구축하였다.

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

2.1.2 대규모 자동차 음성 코퍼스

2차년도(2002. 5. 1 ~ 2003. 4. 30) 자동차 음성 코퍼스 구축 계획은 300명 화자, 5채널 동시 수집, 1인당 100 토큰 규모였으나, 자동차 환경에서의 음성 인식에 대한 관심과 요구가 증대되면서 업체의 요구가 많아 그 규모를 400명 화자, 8채널 동시 수집, 1인당 200여 토큰 규모로 확대하였다. 다채널 대규모 자동차 음성 코퍼스 수집을 위한 발성 환경과 목록은 다음과 같다.

- 설정 환경 : 2환경
  - 환경 1 (50%) : 시내 주행(30~60km/h)
  - 환경 2 (50%) : 고속도로 주행(70~90km/h)
  - 공통 환경 : 맑은 날씨/아스팔트 노면/창문 CLOSE
- 발성 목록 : 총 2,066토큰
  - 1인당 발성량 : 206~207토큰
  - 단독숫자 및 사연숫자 : 900종
  - 다이얼링 명령어 : 63종
  - 카 오디오 명령어 및 관련 단어 : 306종
  - 자동차 컨트롤 스위치 명령어 : 120종
  - 네비게이션 명령어 : 57종
  - PDA 명령어 : 121종
  - 지명(군, 구 단위 이상) : 446종
  - 주요 도로명(고속국도 이상) : 53종

2.2 수출 지원을 위한 외국어 음성 코퍼스

수출 지원을 외국어 음성 코퍼스로 1차년도에는 중국어 음성 코퍼스를 구축하였고, 2차년도에는 대상 언어를 확대하여 영어, 스페인어 음성 코퍼스를 구축하였다.

2.2.1 중국어 음성 코퍼스

중국어 음성 코퍼스의 발성 목록은 421개의 음절, PBW 단어, 사연숫자, 날짜 관련 단어를 포함하는 2,648개의 단어와 400개의 문장으로 구성되어 있다.

음성 코퍼스 수집을 위해 연변대학 지역협력 사이트를 이용하여 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.

2.2.2 영어 음성 코퍼스

영어 음성 코퍼스의 발성목록은 단독숫자, 예/아니오, 알파벳, 화폐단위, 내장 명령어, 특정날짜 시간표현, 응용어, 요일, 비밀번호, 전화번호, 주 및 도시이름, 신용카드 번호를 포함하는 총 1,586개의 단어로 구성되어 있다.

미국 태생의 영어를 모국어로 사용하는 성인 남녀 총 400명의 화자를 대상으로 하여 미국 현지 녹음 수록하였다. 1인당 140~142단어를 발성하였고 사무실 환경에서 Sennheiser m@b40 마이크를 사용하여 PC를 통해 수집되었다.

2.2.3 스페인어 음성 코퍼스

스페인어의 경우 수출 시장을 감안하여 본토 스페인어를 대상으로 하지 않고, 미국 내에서 거주하는 히스페닉계 사람들이 사용하는 스페인어(히스페닉 스페니쉬)를 대상으로 하여 음성 코퍼스를 구축하였다.

발성목록은 단독숫자, 알파벳, 제어명령어, 화폐 단위, 특정 날짜 시간 표현, 응용어, 날짜, 비밀번호, 전화번호, 신용카드 번호를 포함하는 1,230개의 단어와 5,670개의 문장으로 구성되어 있다.

음성 코퍼스 수집을 위해 스페인어를 구사하는 총 300명의 화자(남자 154명, 여자 146명)를 대상으로 미국 서남부 현지에서 녹음을 하였으며, 1인당 120~123 토큰을 발성하였다. 영어음성 코퍼스와 마찬가지로 사무실 환경에서 Sennheiser m@b40 마이크를 사용하여 PC를 통해 수집되었다.

2.3 산업응용 기반기술 기초연구용 음성 코퍼스

2.3.1 클린 스피치 PRW 음성 코퍼스

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

센터의 지역협력 사이트를 활용하여 전국적으로 500명 화자의 음성 데이터를 방음실에서 수집하였으며 남녀 성비는 1:1이고, 1인당 417 ~ 418단어를 발성하였다. 또한 수집된 음성 데이터 전량에 대하여 음운 레이블링 기준(센터 권고안)에 의해 지역 협력 사이트에서 레이블링 전문 인력에 의해 레이블링 검증 및 수정 작업을 진행 중이다.

2.3.2 클린 스피치 낭독 음성 코퍼스

1차년도에 수집된 클린스피치를 단어에서 문장으로 확대 구축하기 위해 그 대상이 되는 발성목록의 설계 과정은 다음과 같고, 구성된 총 20,000여 문장을 200명의 화자를 대상으로 방음실 환경에서 수집하였다.

(1) 발성목록 선정을 위한 말 뭉치의 형태 통계 분석

발성 목록 선정을 위한 모집단으로는 21세기 세종계획 형태소분석 균형 말뭉치 1,000만어절[2]을 사용하였다. 먼저 형태소를 최소 단위로 말뭉치를 분석하

고 분석된 말뭉치의 형태소를 통계 처리하여 각 형태소 유형의 빈도를 조사하였다.

(2) 형태 통계 결과의 정렬

형태소의 통계 분석 결과는 고빈도의 형태소 토큰에서 저빈도의 형태소의 토큰 순으로 정렬하여 형태소, 태그, 상대빈도, 누적빈도, 누적상대빈도, 찌프상수를 분석하였다.

(3) 문장 색인

형태소 분석 말뭉치에서 나열된 어절은 문장으로 복구하고 형태소 분석 부분은 형태 통계에서 한 문장이 갖는 최저 형태소 빈도를 찾아 문장 색인에 사용하였다.

(4) 기존 발성 문장과 비교 선정

색인된 문장은 1차년도 Dictation 낭독 음성 코퍼스 발성목록과 비교하여 기존 발성 목록에 포함되는 문장을 삭제하였다.

(5) 발성 목록 추출

형태소 최저 빈도수로 색인된 문장에서 적당량의 문장을 추출하기 위해서 형태소 최저 빈도수의 임계치가 구해져야 한다. 이 때, 6어절 이상 25어절 이하의 문장만이 누적 문장수의 계산에 사용되었다. 최종 발성 목록은 형태소 최저 빈도수가 826 이상인 20,217 문장에 대해서 수작업에 의한 수정 및 검증을 하였으며, 이전 연구에 의해 구성된 PBS 589문장을 추가하였다.

2.4 산업 응용을 위한 운율 합성용 음성 코퍼스

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 4,392문장이며 이 문장에 포함된 Triphone의 총 종류수는 모집단과 같이 18,025 종류이다.

남, 녀 전문 성우 각 1인이 방음실에서 발성하였다. 마이크는 Rode NT-2를 사용하였으며 EGG 신호도 동시에 수집되었다. 수집된 남, 녀 음성데이터는 모두 음운 및 K-ToBI (Korean-Tone and Break Index) 기준 (Ver 3.1)을 적용하여 운율 레이블링을 실시하였고, 현재 레이블링 결과에 대한 검증을 진행하고 있다.

2.5 Dictation용 낭독 음성 코퍼스

2.5.1 PC 환경 낭독 문장 코퍼스

발성 목록 선정을 위한 모집단으로는 KAIST에서 구축된 4,300만 어절의 KAIST Corpus를 사용하여 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. 본 코퍼스에서는 발성 목록 선정을 위해 고빈도 5,000어절, 8,000어절 및 10,000어절을 발성 목록 선정을 위한 대상어휘로 선정하였다.

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 1인당 평균 104.17문장을 발성할 수 있도록 200개의 발성 세트로 재구성하여 남, 녀 각 200명의 화자에게 음성 데이터를 수집하였으며 1인당 발성량은 104 ~ 105문장이다. 데이터는 사무실 환경에서 PC의 사운드카드와 Andrea ANC 750 마이크를 이용하여 수집되었다.

2.5.2 PC 환경 낭독 문장 코퍼스의 확장

대어휘 연속 음성 인식을 위한 문장 음성 코퍼스를 구축하기 위해 산업응용 기초 연구용 코퍼스의 발성 목록과 동일한 총 20,000여 문장을 발성목록으로 사용하여 총 400명의 화자를 추가하여 PC환경 낭독 문장 음성 코퍼스를 확장하였다.

2.6 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
- PBW : 452 종
- 명령 및 지시어 : 400종
- 단독 숫자 : 41종

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

---

2) 문화관광부, 국립국어연구원(2002). 21세기 "세종계획 2001년도 국어 기초자료 구축 분과 연구 결과 보고서" http://www.sejong. or.kr

하여 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 요인 중 하나이다. 따라서 센터에서는 이러한 다양한 변인에 따른 시험용 코퍼스의 구축을 위해 1차년도에는 마이크의 종류에 특성변화 시험용 음성 코퍼스를 구축하였고, 2차년도에는 마이크로폰의 거리에 따른 영향을 분석하기 위한 음성 코퍼스를 구축하였다.

수집절차는 방음실에서 고성능 마이크로폰으로 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 총 8종을 대상으로 방음실 환경에서 데이터를 수집하였다. 2차년도에는 마이크의 거리를 5cm, 10cm, 20cm, 50cm, 100cm로 달리하여 AKG C400-BL, SENNHEISER e-825s 두 종류의 마이크를 대상으로 하여 데이터를 수집하였다.

2.8 기기내장형 음성 코퍼스

최근 다양한 음성정보기술이 실생활에 적용되기에 이르렀고, 차세대 사용자 인터페이스 수단으로 부각되면서 완구, 로봇, PDA, 홈오토메이션과 같은 다양한 임베디드용 음성 인식 어플리케이션이 개발되고 있다. 이처럼 다양한 기기내장형 음성 인식용 코퍼스 구축을 위한 발성목록은 다음과 같다.

- PDA 공통 명령어 : 총 12종
- 단독 숫자 : 12종
- PRW(Phonetically Rich Words) : 4,175종

USB-DSP 임베디드용 음성 수집 툴킷을 이용하여 총 300명의 화자를 대상으로 기기 내장형 electret 콘덴서 마이크를 통해 수집하였으며 1인당 107 ~ 108단어를 발성하였다.

2.9 숫자음 코퍼스

1차년도에는 PC 환경 숫자음 500명분을 구축하였고, 2차년도에는 그 대상 및 범위를 확대하여 PC 환경 및 전화망 환경에서 음성 인식용 숫자음 코퍼스를 구축하였다. 숫자음 코퍼스의 확장을 위한 발성목록은 총 25,000종의 2~3 음절로 이루어진 단위 숫자로 구성되어 있다. 총 500명의 화자를 대상으로 PC환경에서는 Sennheiser E 835S 다이나믹 마이크를 사용하였고, 전화망 환경에서는 인텔 Dialogic보드를 이용하여 유선, 무선, 셀룰라, PCS 환경 분포를 고려하여 수집하였다.

2.10 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
- Web TV의 제어명령어 인식용 음성 코퍼스

III. 3차년도 음성 코퍼스 수집 계획

센터에서는 차기연도 음성 코퍼스 구축을 위한 계획을 작성하기 위하여 음성정보기술 관련 전문가들을 대상으로 수요조사를 실시하였다. 차기연도의 음성 코퍼스의 구축계획은 조사된 결과 및 R&D Roadmap 등을 참고하여 결정하였으며 자세한 내용은 표 1과 같다.

표 1. 3차년도 음성 코퍼스 구축 계획

명칭	규격	규모
수출지원을 위한 외국어 음성 코퍼스	• 인식용 외국어 코퍼스 - 영어 : PRW 및 단문 - 중국어 : 단어 및 문장 확장	영어:400명분 중국어:300명분
자동차 음성 코퍼스	• 실차 환경 음성 코퍼스 확장 - 발성내용, 수록조건, 발성자 수의 확장	5채널 400명분
산업응용을 위한 화자인증 음성코퍼스	• 시차별	30명분
멀티모달 음성 코퍼스	• 음성 및 입술 영상 코퍼스	200명분
음성인식 성능 평가용 음성 코퍼스	• 실제 환경에서의 인식 성능 평가용 음성 코퍼스 • AURORA 2.0 수준	300명분
복지 응용을 위한 음성 코퍼스	• 복지 응용을 위한 장애자 및 노인 음성 프로토타입	100명분
모의 환경 음성 코퍼스	• 자동차 환경	다채널 500명분
대화체 인식용 음성 코퍼스	• 수집 분야, 수집 방법 및 정보표기법 기초검토 • 소규모 프로토타입 시험 제작	
한국인 외국어 음성 코퍼스	• 소요 타당성 및 스펙 예비 검토	
대화체 합성용 음성 코퍼스	• 소요 타당성 및 스펙 예비 검토	
고소음 음성 코퍼스	• 소요 타당성 및 스펙 예비 검토	
감정 음성 합성 연구용 코퍼스	• 소요 타당성 및 스펙 예비 검토	

IV. 결론

본 논문에서는 SiTEC에서 구축중인 음성코퍼스의 현황과 향후 계획에 관하여 보고하였다. 자동차 내 소음 및 음성 코퍼스, 다양한 외국어 음성 코퍼스, 방음실 환경에서의 인식 및 합성 연구용 코퍼스, 아동용 음성 코퍼스, Dictation용 음성코퍼스 등의 구축 내용이 소개되었다. 센터에서는 향후 현재 구축된 음성 코퍼스의 내용과 양을 지속적으로 보완하고 확장할 계획이며, 음성 코퍼스의 구축 내용과 방향에 대한 관련

연구자들의 많은 참여와 의견을 기대하고 있다.

참고 문헌

[1] 이용주, "음성언어코퍼스,"한국정보과학회지, 1998.2
[2] 원광대학교 음성정보기술산업지원센터 홈페이지 :
http://www.sitec.or.kr