SiTEC의 STiLL 관련 음성 코퍼스의 구축 현황

김영일* 김봉완* 최대림* 이광현* 정은순* 이용주**
* 원광대학교 음성정보기술산업지원센터
** 원광대학교 전기전자 및 정보공학부

Creation of Speech Corpora for STiLL at SiTEC

YoungIl Kim*, BongWan Kim*, DaeLim Choi*, KwangHyun Lee*, EunSoon Jeong*, YongJu Lee**
* Speech Information Technology & Industry Promotion Center
** Department of Electrical, Electronic and Information Engineering, Wonkwang University

*{yikim, bwkim, dlchoi, khlee, true}@sitec.or.kr, **yjlee@wonkwang.ac.kr

Abstract

As language learning that utilizes speech and information processing technology is getting popular. Speech Information Technology & Promotion Center (SiTEC) has created and is distributing speech corpora for STiLL in order to support basic research and development of products. We will introduce the corpus for Korean and those for English which we have created and are distributing.

Ⅰ. 서론

음성정보기술산업지원센터(이하 센터)에서는 기존에 구축된 음성 코퍼스의 내용과 양을 지속적으로 보완하고 확장하는 동시에, 음성정보기술의 연구 및 상품 개발에 필요로 하는 새로운 코퍼스를 조사하여 구축해 나가고 있다. 2001년도에 설립된 이후 현재까지 총 34종의 음성 코퍼스를 구축하여 배포하고 있다[1].

최근 국내 및 국외 언어교육 시장이 활성화되어가고 있으며, 음성인식과 음성합성을 접목시킨 제품들의 출시가 늘어나고 있다. 이를 위해 센터에서는 음성정보기술 관련 산업의 응용 범위 확대를 위하여 언어교육 분야에 사용되어질 수 있는 음성 코퍼스를 구축하여 언어교육 연구 및 실제 응용 가능한 어플리케이션 개발을 지원하고자 한다.

본 논문에서는 센터에서 보급하고 있는 언어 교육 과 관련된 음성 코퍼스를 소개한다.

Ⅱ. STiLL 관련 음성코퍼스 구축 현황

센터의 STiLL 관련 음성 코퍼스는 크게 한국어 교육과 영어 교육 분야로 구분할 수 있다. 특히, 영어 교육과 관련해서 산업체의 해외 시장 진출을 지원하기 위하여 중국인의 영어 음성 코퍼스와 일본인의 영어 음성 코퍼스를 구축 중에 있다.

본 논문에서는 센터에서 보급하고 있거나 보급 예정인 한국어 교육을 위한 '외국인의 한국어 음성 코퍼스'와 영어 교육을 위한 '중국인의 영어 음성코퍼스', '일본인의 영어 음성 코퍼스', '한국인의 영어 음성 코퍼스'에 대해 기술한다.

1. 외국인의 한국어 음성 코퍼스

외국인들이 한국어를 발화할 때 나타나는 특성에 관한 연구와 한국어 학습을 위한 어플리케이션 개발을 지원하기 위하여 외국인이 발성한 한국어 음성 코퍼스를

구축하였다[2]. 외국인의 한국어 교육을 위해서는 한국어 학습자의 음성뿐만 아니라, 교육용 샘플이 될 수 있는 한국인의 음성도 필요하다고 판단되어, 외국인과 한국인의 음성을 수집하였다.

발성목록은 기존 한국어 교육 경험을 토대로 외국인 학습자들이 주로 나타내는 오류를 정의하고 이에 적합한 문장을 새롭게 설계하였다. 정의된 오류의 유형을 살펴보면, 중국어권, 영어권, 일본어권 화자들 모두 경음, 격음, 평음을 구분하지 못하고, 이중모음과 'ㅢ', 'ㅚ', 'ㅟ'를 정확하게 발성하지 못하는 경향을 보인다. 그리고 언어권에 따라서는 대표적으로 다음과 같은 특징들을 보인다.

○ 중국어권 화자
- 음절 초 'ㄹ'을 [l]로 발음.
- 유음 받침을 권설음 [r]로 발음.
- 특정 분절음 연쇄를 발음하지 못함.
(은행->으이항, 년->니엔)

○ 영어권 화자
- 음절 초 'ㄹ'을 [r]로 발음.

○ 일본어권
- 받침 발음을 정확히 못함.
(김치->기무치, 밥도->받또, 딸기->따루기)
- 이중모음을 단모음으로 발음함.

위 특징들을 반영한 69개 어절로 구성된 단문 10개와 대화문 20개, 그리고 초급 학습자들도 이해할 수 있는 쉬운 단어 88개로 구성된 발성목록을 설계하였다. 한 명의 화자가 단문과 대화문을 1회씩 발성하고, 단어를 2회 발성하였으며, 대화문은 두 명의 화자가 실제 서로 대화하도록 하여 녹음하였다.

발성화자는 한국어 교육 기관에서 한국어를 학습하고 있는 20, 30대 외국인 180명과 한국인 20명으로 구성되어 있다. 발성목록 설계에서 고려한 것처럼, 외국인 학습자의 경우에 언어권에 따라 특성이 다르므로 이를 반영하기 위하여 영어권, 중국어권, 일본어권 화자로 구분하여 수집하였다. 또한 같은 언어권 내에서도 출신 지역에 따라 다른 양상을 보이므로 영어권 학습자의 경우에는 북미, 중국어권 학습자는 북경, 일본어권 학습자는 동경 지역 출신으로 한정하였다. 그리고 학습자의 언어 숙달도에 따른 특성을 살펴보기 위해 사전 인터뷰를 통해 학습자를 크게 초급, 중급, 고급의 세 그룹으로 분류하였다.

외국인 학습자의 언어권과 숙달도에 따른 인원구성은 [표 1]과 같다.

숙달도 / 언어권	초급		중급		고급		합계
	남	여	남	여	남	여	
중국어권(북경지역)	11	11	12	10	10	11	65
일본어권(동경지역)	9	9	7	8	6	5	44
영어권(북미지역)	10	11	10	13	12	15	71
합계	30	31	29	31	28	31	180

표 1 인원구성 (단위: 명)

방음실에서 HMD 25-1 마이크와 DAT 레코더를 이용하여 16kHz, 16Bit, Windows PCM 포맷으로 녹음하였다. 수집된 음성 데이터는 총 40,829개이며, 용량은 2GB이다.

2. 수출 지원을 위한 외국인의 영어 음성 코퍼스

산업체의 해외 시장 진출을 지원하기 위하여 영어 교육에 초점을 맞추어 외국인의 영어 음성 코퍼스를 구축하여 보급하고 있고 일부는 구축 중에 있다.

2.1 발성목록

학습자들이 외국어를 배우는데 있어서 어려운 분야 중에 하나가 발음이다. 따라서 영어 교육을 위한 음성 코퍼스의 발성목록은 영어 학습 과정에서 나타나는 다양한 발음 현상을 폭넓게 포함하는 목록을 사용하는 것이 바람직하다. 이를 위해 단어 선정과 문장 선정 기준을 마련하여 목록을 설계하였다.

단어 선정을 위한 기준은 다음과 같다.

○ 여러 사용자의 발음 교정을 목적으로 하므로 사용 빈도가 높은 초등, 중등의 기초어휘를 포함하도록 한다.

○ 명사의 단수, 동사의 원형뿐만 아니라, 복수형과 활용형을 포함하도록 한다.

○ 다양한 접미사 및 접두사로 이루어진 단어를 포함하도록 한다.

○ 음소 대조를 통한 학습을 위해 최소 대립쌍을 포함하며, 다섯째, 영어의 어두 및 어말 자음군을 포함하도록 한다.

문장을 사용하는 이유는 음소와 음소 사이에 일어날 수 있는 변이를 단어 수준에서 문장 수준으로 확장해서 파악하기 위함이다. 이에 따라 문장 내에서 나타나는 연음현상을 정의하고 발생 가능한 오류에 따라 문장을 선정하였다.

단어목록과 문장목록을 만들기 위하여 사용된 자료는

다음과 같다.

- 시사 영어 사전의 초등, 중등 기본 어휘
  - Oxford의 Mr. Bug's Phonics와 Up and Away Phonics
  - Cambridge University Press의 Elements of Pronunciation의 어두, 어말 자음군 단어
  - John Trim의 English Pronunciation Illustrated의 최소 대립쌍 단어

최종 선정된 단어는 총 2,951개이며, 문장은 총 125개이다. 단어는 20개 세트로 나누어 한 명의 화자가 148개 정도를 발성하도록 하였으며, 문장은 10개 세트로 나누어 한 화자가 12개 정도를 발성하도록 하였다.

2.2 전사

영어 음성 데이터 전사를 위해서는 별도의 전사 기준이 필요하다. 이를 위해 The CMU Pronouncing Dictionary[3]와 Klatt Symbols[4]을 참조하여 총 53개의 음소를 정의하였다. 정의된 음소 기호에 의해 음성 데이터 전량을 수동으로 전사하여 배포하고 있다.

2.3 종류

센터에서는 수출 지원을 위한 외국인의 영어 음성 코퍼스로 '중국인의 영어 음성 코퍼스'를 구축하여 보급하고 있으며, 이를 확대하기 위하여 현재 '일본인의 영어 음성 코퍼스'를 만들고 있다.

중국인의 영어 음성 코퍼스[5]는 중국 북경에 거주하며 북경어를 모국어로 사용하는 남자 100명, 여자 100명으로 구성되어 있다. 연령에 따라 20대 65%와 30대 35% 비율로 구성되어 있다. 데이터는 조용한 사무실 환경에서 Labtec Axis-301 마이크와 PC를 이용하여 16kHz, 16Bit, Windows PCM 포맷으로 녹음되었다. 전체 데이터에 대해서 수동으로 발음전사를 수행하여 배포하고 있으며, 음성 데이터는 총 31,990개이고 용량은 1.5GB 이다.

일본인의 영어 음성 코퍼스는 현재 구축 중이며, 일본어를 모국어로 사용하는 20, 30대 화자 200명의 영어 음성을 수집하고 있다. 발성 목록은 2.1절에서 기술한 단어 목록과 문장 목록을 사용하고 있다.

조용한 사무실 환경에서 Labtec Axis-301 마이크와 PC를 이용하여 16kHz, 16Bit, Windows PCM 포맷으로 수집하고 있으며, 발음 전사를 완료한 후 보급할 예정이다.

3. 한국인의 영어 음성 코퍼스

한국인의 영어 음성 코퍼스[6]는 한국학술진흥재단(과제번호 2002-042-A00035)의 연구비 지원으로, 이석재(연세대), 이용주(원광대), 이숙향(원광대), 강석근(원광대)에 의해 제작되었으며, 센터에서 보급하고 있다.

다양한 대상 기준에 따른 한국인의 영어 발음 음성을 수집하여 실험음성학과 음운론, 그리고 영어 교육을 위한 기초적 자료와 연구 대상을 제공하기 위하여 구축되었다.

이를 위해 한국인과 원어민이 발성한 음성을 수집하였으며, 발성화자는 초등, 중등, 일반인을 대상으로 하여 총 342명이 참여하였다. 지역별로는 서울, 경기, 충청, 전라, 경상, 강원, 제주 지역으로 구분하여 수집되었다.

구분		남자	여자	합계
한국인 (327)	일반	56	57	113
	중등	49	53	102
	초등	56	56	112
외국인 (15)	일반	6	6	12
	초등	2	1	3
합계		169	173	342

표 2 초등, 중등, 일반인 구성(단위:명)

	남자			여자		
	일반	중등	초등	일반	중등	초등
서울	8	8	8	9	8	8
경기	8	8	8	8	8	8
충청	8	4	8	8	5	8
전라	8	7	8	8	7	8
경상	8	6	8	8	8	8
강원	8	8	8	8	8	8
제주	8	8	8	8	9	8
외국인	6		2	6		1

표 3 지역별 인원 구성 (단위: 명)

발성목록은 영어의 분절적인 면과 초분절적인 요소의 특징을 복합적으로 나타낼 수 있는 어휘, 문장 그리고 이야기 읽기로 설계되었다. 또한 동일 화자의 모국어와 외국어를 발화할 때의 음성 비교를 위하여 기본적인 한국어 단어도 포함되었다. 발성목록은 총 6개 세트로 그 내용은, 한국어 기본단어 98개, 한국어 바람과 해님 이야기, 영어 자음 및 모음 관찰용 단어 64개, 영어 어휘, 영어 문장 36개, 영어 바람과 해님 이야기로 구성되어 있다. 영어 어휘는 난이도에 따라 초등 435개, 중등 956개, 일반인 1,125개로 구분되어 있으며, 한 명의 화자가 6개 세트를 모두 발성하였다.

방음실 환경에서 Sennheiser HMD25-1과 Shure SM10A 마이크와 PC를 이용하여 16kHz, 16Bit, Windows PCM 포맷으로 수집되었다. 코퍼스의 음성

파일은 총 328,313개이며, 용량은 14GB이다.

Ⅲ. 결론

본 논문에서는 음성정보기술산업지원센터(SiTEC)에서 구축하여 보급 중인 STiLL 관련 음성 코퍼스인 외국인의 한국어 음성 코퍼스, 중국인의 영어 음성 코퍼스, 일본인의 영어 음성 코퍼스, 한국인의 영어 음성 코퍼스에 대해서 소개하였다. 이러한 음성 코퍼스들이 국내의 STiLL관련 연구자들에게 널리 사용될 수 있기를 바라며, 이러한 일련의 연구들이 대규모로 기획되기를 희망한다.

참고문헌

[1] 원광대학교 음성정보기술산업지원센터 홈페이지, http://www.sitec.or.kr

[2] 신지영, 이석재, "외국인의 한국어 음성 코퍼스-발성목록 설계 및 데이터 수집" 용역 보고서, 음성정보기술산업지원센터, 2005.

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[4] ALLEN, J.- HUNNICUTT, M.S.- KLATT, D.H. (with R.C. ARMSTRONG and D. PISONI) (1987) From Text to Speech: The MITalk System. *Cambridge: Cambridge University Press* (Cambridge Studies in Speech Science and Communication). [App. B "Klatt symbols"]

[5] 휴먼미디어테크, "중국인의 영어 음성 코퍼스 구축" 용역 보고서, 음성정보기술산업지원센터, 2005.

[6] 이석재(연세대), 이숙향(원광대), 강석근(원광대), 이용주(원광대), "한국인의 영어 음성 코퍼스 설계 및 구축", 말소리, 제46호, pp.159-174, 2003.