말소리와 음성과학 제7권 제1호 (2015.03.31)
ISSN 2005-8063, pp. 139~150
www.speechsciences.or.kr
http://dx.doi.org/10.13064/KSSS.2015.7.1.139

한국인 표준 음성 DB 구축

Developing a Korean Standard Speech DB

신 지 영[1] · 장 혜 진[2] · 강 연 민[3] · 김 경 화[4]
Shin, Jiyoung · Jang, Hyejin · Kang, Younmin · Kim, Kyung-Wha

ABSTRACT

The data accumulated in this database will be used to develop a speaker identification system. This may also be applied towards, but not limited to, fields of phonetic studies, sociolinguistics, and language pathology. We plan to supplement the large-scale speech corpus next year, in terms of research methodology and content, to better answer the needs of diverse fields. The purpose of this study is to develop a speech corpus for standard Korean speech. For the samples to viably represent the state of spoken Korean, demographic factors were considered to modulate a balanced spread of age, gender, and dialects. Nine separate regional dialects were categorized, and five age groups were established from individuals in their 20s to 60s. A speech-sample collection protocol was developed for the purpose of this study where each speaker performs five tasks: two reading tasks, two semi-spontaneous speech tasks, and one spontaneous speech task. This particular configuration of sample data collection accommodates gathering of rich and well-balanced speech-samples across various speech types, and is expected to improve the utility of the speech corpus developed in this study. Samples from 639 individuals were collected using the protocol. Speech samples were collected also from other sources, for a combined total of samples from 1,012 individuals.

Keywords: Korean standard speech corpus, read speech, spontaneous speech, speaker identification

1. 서론

언어 연구를 위해 구축된 코퍼스는 대표성과 균형성으로 평가할 수 있다. 코퍼스의 대표성은 모집단 선정이나 표본 텍스트의 선정 및 전체적인 양과 크기에 따라 획득되고, 균형성은 코퍼스를 구성하는 텍스트 요소들 사이의 균형과 가중치에 따라 획득된다[1]. 한국어 음성 코퍼스가 여러 연구 기관에서 각각의 목적에 따라 다양하게 구축되고 있으나, 대표성과 균형성을 모두 갖춘 것을 찾아보기는 어렵다. 음성학적 목적에서 한국인의 음성 특징을 알아내기 위한 자료로서는 물론, 방언학, 사회 언어학 등 한국어의 여러 면모를 살피는 연구에서 대표성과 균형성을 갖춘 음성 코퍼스의 구축은 필수적이다. 비단 언어학적 연구뿐 아니라 다양한 응용 분야, 즉 음성 공학이나 언어 병리학 등의 연구를 위한 기초 자료로서도 음성 코퍼스가 반드시 필요하다.

이 연구는 대검찰청에서 발주한 '용의자 음성식별을 위한 한국인 음성 데이터베이스 수집 및 음성 자동분석 시스템 개발(2014.5.~2014.11.)' 과제의 수행을 위해 구축하는 한국인 표준 음성 DB의 구축 방법을 보고하고, 구체적인 자료 수집 방법을 소개하는 데 목적이 있다.[5] 이 과제에서 구축하는 한국인

1) 고려대학교, shinjy@korea.ac.kr
2) 고려대학교, jina49@korea.ac.kr, 교신저자
3) 고려대학교, flour@korea.ac.kr
4) 대검찰청, savoix@spo.go.kr

이 논문은 2014년 대검찰청 연구 용역의 지원으로 수행되었습니다. (과제명: 용의자 음성식별을 위한 한국인 음성 데이터베이스 수집 및 음성 자동분석 시스템 개발)

접수일자: 2015년 2월 07일
수정일자: 2015년 3월 11일
게재결정: 2015년 3월 11일

5) 해당 과제는 3년간 총 3,000명 이상의 한국인 음성 데이터베이스를 수집하는 것을 목표로 기획되었다. 본 연구는 이 가운데 1차 연도 과제 수행 결과를 보고하는 것으로, 전체 수집 인원의 1/3에 해당하는 1,000명의 자료 수집 과정과

표준 음성 DB는 인구 통계학적 정보에 기반하여 화자 집단을 설정함으로써 한국인의 음성에 대한 대표성을 갖추고, 같은 화자의 여러 형태 발화를 수집함으로써 균형성을 갖추고자 하였다. 이 한국인 표준 음성 DB는 한국어의 하위 방언권 전체를 범위로 하여, 20대에서 60대에 이르는 다양한 연령층의 남녀 화자를 대상으로 설정하였다. 또한 모든 화자에 대하여 낭독 발화와 자유 발화를 포괄하는 다양한 유형의 발화 음성을 수집하기 위하여 한 화자에게서 총 5가지 유형의 발화 자료를 단계별로 수집함으로써 규모는 물론 내용에 있어서도 대표성과 균형성을 확보하고자 하였다.

2. 국내외 음성 코퍼스 현황

2.1 국내 음성 코퍼스

국내 음성 코퍼스는 그 규모나 연구의 축적 측면에서 높은 수준에 이르렀다고 하기 어렵다. 현재 연구 목적으로 활용할 수 있도록 공개되어 있는 음성 코퍼스로는 국립국어원에서 구축한 서울말 낭독체 발화 말뭉치, 음성정보기술산업지원센터(SiTEC)에서 구축한 외국인의 한국어 발화 음성 DB 등이 있다. 서울말 낭독체 발화 말뭉치는 20대~50대 이상 남녀 화자 총 120명의 자료를 수집한 것이다. 이 코퍼스는 연령을 세분화하여 자료를 수집하였고, 음성 자료와 대본이 충실하게 구비되어 있다는 장점이 있다. 그러나 '서울말', '낭독체' 발화라는 점에서 지역과 발화 유형이 한정적이다.

외국인의 한국어 발화 음성 DB에서는 외국인과 비교를 위해 함께 수집한 한국인의 음성을 활용할 수 있다. 그러나 화자가 20-30대 서울말 화자에 한정되어 있으며, 화자 수도 20명으로 소규모라는 한계가 있다.

한편 국립국어원의 발간물 목록을 살펴보면 각 지역 지역어 전사 보고서, 국어 음성 분석 연구 등과 같은 연구에서 음성 수집이 있었을 것으로 생각되나 존재 여부를 확인하기가 어렵다. 이 외에도 한국학중앙연구원에서 구축한 '한국학중앙연구원 한국 방언 자료집'에서 방언 조사를 통해 채록한 방언 음성을 SiTEC에서 디지털화한 DB가 있다. 이 자료는 음성 파일 스트리밍 형태로 자료를 제공하고 있는데, 자료에 대한 정보로는 구술자와 채록자의 이름과 채록 지역만 확인 가능하다.[6]

각 대학 및 연구소 등에서도 연구 목적으로 구축한 음성 코퍼스가 있다. 이 가운데 고려대학교 민족문화연구원 음성언어센터에서 보유하고 있는 음성 DB는 음성 자료는 물론, 여러 층위의 전사 및 분석 자료가 구비되어 있는 음성 코퍼스이다. 대표적으로 자유 발화 DB가 있다. 이 DB는 자유로운 대화 발화를 수집한 것으로서, 연령에 따라 성인(20-30대)과 청소년(중2, 고2), 초등학생(초등학교 1·3·6학년), 아동(3~8세) DB로 구분되어 있다. 아동 DB의 경우 아동 1명과 통제자 1명이 놀이와 책을 통해 대화한 것이고, 나머지 자유 발화 DB는 모두 해당 연령의 화자들이 3인 1조로 자유롭게 대화한 것을 녹음하여 구축한 것이다. 자유 발화 DB의 규모는 <표 1>에 제시한 바와 같다.

표 1. 고려대학교 민족문화연구원 음성언어센터의 자유 발화 DB 규모

	아동	초등학생	청소년	성인
화자 수	53명	54명	39명	57명
어절 수	61,334	65,825	119,546	174,409
분량	25시간	9시간	13시간	19시간

이 가운데 성인 자유 발화 자료는 철자 전사와 함께 음운 전사가 이루어져 있고, 운율 레이블링이 수행된 상태여서 다양한 목적의 음성 언어 연구에 유용하게 사용할 수 있다. 그러나 서울 방언만을 대상으로 하였으며, 성인의 경우 20-30대 화자의 자료만이 포함되어 있다는 점에서 한계가 있다.

국내 음성 코퍼스를 검토해 본 결과 기존의 코퍼스들은 연구 목적으로 활용하기에는 규모나 접근성 등의 측면에서 한계가 있다는 것을 알 수 있다. 또한 서울 지역 20~30대 화자의음성에만 편중되어 있어 한국어 음성의 전반적인 모습을 관찰할 수 있는 자료는 찾아보기 어렵다.

2.2 국외 음성 코퍼스

국외의 경우 다양한 목적과 형태의 음성 코퍼스가 구축되고 있다. 먼저 벅아이 코퍼스(Buckeye corpus)는 자유 발화 음성 코퍼스로 최근 주목받는 코퍼스이다. 벅아이 코퍼스[7]는 성별과 연령을 균등하게 배분한 40명의 화자에 대해 각각 1시간가량 면접을 실시하여 녹음을 수행하였다. 백인, 오하이오 콜럼버스의 거주자, 상위 노동자~중산층 계급으로 화자를 한정하고, 일상적 주제의 자유 발화를 수집한 공개 코퍼스로, 총 30만 단어 규모이다. 음성 녹음 방법은 인터뷰 형식이었으며, 조사자와 피험자가 대화를 하는 상황에서 피험자에게만 헤드셋 마이크를 착용하게 하여 피험자의 음성만 수집하는 방식으로 진행되었다.[8]

캠브리지 대학에서 수행한 DyViS(Dynamic Variability in Speech)[9]는 범죄 수사 음성학 연구를 위해 구축된 음성 코퍼

결과를 제시한다.

6) http://yoksa.aks.ac.kr/

7) http://buckeyecorpus.osu.edu/php/corpusInfo.php

8) 이와 같은 코퍼스 구축 방법을 한국어에 도입하여 한국의 벅아이 코퍼스를 구축하려는 시도가 있다. 이 코퍼스는 벅아이 코퍼스의 구축 방법 및 규모를 따라 현재 구축 중에 있다[3][4]. 한국의 벅아이 코퍼스는 10대~40대로 비교적 넓은 연령대를 포괄하려 하였으나, 지역은 서울 방언으로 한정하였으며, 전체 화자의 수가 40명으로 그 규모가 작은 편이다.

스이다. DyViS는 화자 변별에 주된 목적을 둔 코퍼스로서, 큰 집단 안의 화자를 변별하기 위한 '화자 공간'을 현실화하고, 개별 화자의 역동적인 조음-음향 특질을 수량화하는 음향 음성학적 목표를 가진다. 그밖에도 범죄 수사용 음성 전문가 및 그 외 관심 있는 연구자들이 널리 사용할 수 있는 데이터베이스를 표방한다. 화자는 SSBE(Standard Southern British English)를 구사하는 18-25세 남성 100명으로, 경찰 심문, 전화 대화, 뉴스 보고(낭독), 통제 문장(낭독) 등의 다양한 양식으로 발화한 자료를 수집하였다. 이 연구는 개별 화자 특징 및 수사 목적의 화자 식별 연구에 유용하며, 영국 영어의 현시대 표준 발음에 있어서 활용 가능한 최대 규모 자료라는 점에서 의의를 가진다[5].

TIMIT 코퍼스(TIMIT Acoustic-Phonetic Continuous Speech Corpus)[10]는 음향음성학적 지식과 음성 인식 시스템 개발에 데이터를 제공하기 위한 목적을 가지고, MIT(Massachusetts Institute of Technology), SRI(SRI International), TI(Texas Instruments, Inc)의 협력으로 개발된 낭독 음성 코퍼스이다[6]. 이 코퍼스는 미국 영어를 8개 방언권으로 구획하여 총 630명의 음성을 수집한 것으로, 총 6,300개 문장, 5.4시간 규모의 음성 자료와 전사 자료로 구성되어 있다. 각 화자는 10개의 문장을 낭독하였고, 이에 대한 전사는 방언 및 음성적 변이에 대한 포괄성을 고려하여 이루어졌다. 음성 및 전사 자료는 미국 표준기술연구원(NIST)에서 인증하였다. 이 코퍼스는 방언별로 화자를 균등하게 모집하였으며, 방언 구획에 있어서 이주가 잦아 지역적 정체성이 불분명한 집단의 방언을 8번째 방언으로 포함함으로써 단일 지역이라는 전통적인 지역 방언 배경을 갖지 않는 화자 집단의 방언 현상을 포함하였다. 다만, 이 코퍼스는 남성 화자의 비율이 70%로, 성별 불균형을 문제 삼을 수 있다. 또한 모든 화자가 같은 문장 세트를 낭독한 것이 아니라는 특징이 있다.

지금까지 국외 음성 코퍼스를 살펴본 결과 다양한 목적과 수집 방식으로 서로 다른 규모의 코퍼스가 구축되어 있음을 알 수 있었다. 특히 목적에 맞는 음성을 수집하기 위하여 발화의 양식에 있어서 여러 시도를 하고 있음이 드러난다. 다만 지역별 인구 비례를 고려한 국가 단위의 표준 음성 코퍼스는 찾아보기 어렵다. 따라서 대표성과 균형성 면에서 한국어 표준 음성 DB가 될 수 있는 음성 코퍼스를 설계하고 구축하는 것이 필요하다. 본 연구는 대검찰청의 지원을 받아 1년간 수행한 한국어 표준 음성 DB를 염두에 둔 음성 코퍼스 구축의 설계 및 구축 과정을 보고하고자 한다.

3. 연구 내용

3.1 조사 대상

이 연구는 지역, 연령 및 성별 인구 비례에 따라 화자를 선정하여 인구 통계적 대표성이 있는 한국인의 표준 음성 DB 구축의 내용과 방법을 소개한다. 2010년 통계청 인구총조사[11] 결과에 근거하여 지역별, 연령별, 성별 비율을 배분하여 총 1,000명의 음성 자료 수집 계획을 수립하였다.

먼저 지역은 수도권(서울, 인천, 경기), 경남권(부산, 울산, 경남), 경북권(대구, 경북), 전남권(광주, 전남), 전북권(전북), 충남권(대전, 충남), 충북권(충북), 강원권(강원), 제주권(제주)의 총 9개의 하위 방언권으로 나눈다. 전국적인 조사에 있어서, 지역별 인원을 어떠한 비율로 조사할 것인지에 대해서는 두 가지 방안이 가능한데, 첫째는 지역 균등에 따르는 것이고 둘째는 인구 비례에 따르는 것이다. 이 연구에서는 지역별 인구수 차이를 반영하여 수집 인원을 배정하였다. 즉, 모집단이 큰 지역에 높은 비율을 부여하여, 해당 지역의 표본을 많이 수집하는 방식이다. 이처럼 인구 비례 방식을 택한 것은, 범죄 수사 상황에서 수집된 용의자의 방언과 일치할 확률이 높은 음성을 많이 수집하는 것이 용의자의 음성 식별에 유리할 것으로 판단되기 때문이다.

서울과 6대 광역시, 각 도별 인구 중 내국인의 수를 고려하여 이를 방언권별로 통합한 후, 해당 연도에 목표한 수집 인원 1,000명을 기준으로 하여 각 방언권별로 배분하였다. 방언권별 수집에서는 해당 권역에 속한 인구를 대표하는 도시, 즉 대도시를 중심으로 한다. 왜냐하면 화자의 대표성은 수적인 대표성도 말하는 것이기 때문이다. 지역별 녹음 인원은 <표 2>에 제시한 바와 같다. 서울, 인천, 경기 지역을 합한 수도권의 인구가 전체의 48.9%로 가장 많고, 다음으로 부산, 울산, 경남 지역을 합한 경남권의 인구가 15.8%, 대구, 경북 지역을 합한 경북권이 10.5% 등의 순으로 나타난다.

연령은 20대에서 60대까지 총 5개의 연령대로 나눈다. 인구통계에 따른 연령별 인구수와 비율, 그리고 20~60대를 기준으로 한 연령별 인구 비율을 <표 3>에 제시하였다. 전체 인구 가운데 20대에서 60대의 인구는 69%에 해당한다. 20대에서 60대만을 기준으로 연령에 따른 분포를 살펴보면, 40대가 24.8%로 가장 많고, 30대가 23.5%, 20대와 50대가 각각 19.9%와 19.8%, 60대가 12%의 순으로 나타난다.

성별의 경우 <표 4>에 제시한 바와 같이 남성이 총인구의 49.7%, 여성이 50.3%로, 남녀가 대체로 비슷한 비율로 나타난다. 따라서 연령과 방언을 고려한 피험자 집단 내에서 남녀 화자의 비율을 1:1로 수집한다.

9) http://www.ling.cam.ac.uk/dyvis/

10) https://catalog.ldc.upenn.edu/LDC93S1

11) http://kosis.kr/

표 2. 인구 통계를 반영한 지역별 녹음 인원(1,000명당)

		총인구	%	1,000명당
수도권	서울	9,631,482	20.1	489
	인천	2,632,035	5.5	
	경기	11,196,053	23.3	
경남권	부산	3,393,191	7.1	158
	울산	1,071,673	2.2	
	경남	3,119,571	6.5	
경북권	대구	2,431,774	5.1	105
	경북	2,575,370	5.4	
전남권	광주	1,466,143	3.1	67
	전남	1,728,749	3.6	
전북권	전북	1,766,044	3.7	37
충남권	대전	1,490,158	3.1	73
	충남	2,000,473	4.2	
충북권	충북	1,495,984	3.1	31
강원권	강원	1,463,650	3.0	30
제주권	제주	528,411	1.1	11
합계		47,990,761	100.0	1,000

표 3. 인구 통계를 반영한 연령별 녹음 인원(1,000명당)

	총인구	%	20-60대 기준	
			%	1,000명당
0~9세	4,613,747	9.6	-	-
10~19세	6,611,640	13.8	-	-
20~29세	6,594,369	13.7	19.9	199
30~39세	7,794,495	16.2	23.5	235
40~49세	8,204,781	17.1	24.8	248
50~59세	6,564,826	13.7	19.8	198
60~69세	3,994,404	8.3	12.0	120
70~79세	2,650,381	5.5	-	-
80세 이상	962,118	2.0	-	-
합계	47,990,761	100.0	100.0	1,000

표 4. 인구 통계를 반영한 성별 녹음 인원(1,000명당)

	총인구	%	1,000명당
남성	23,840,896	49.7	497
여성	24,149,865	50.3	503
합계	47,990,761	100.0	1,000

방언권별 구획에 따른 연령대별 인구 통계는 <표 5>에 제시한 바와 같다.

표 5. 방언권별 구획에 따른 연령대별 인구 통계(1,000명당)

	1,000명당	20대	30대	40대	50대	60대
수도권	489	97	115	122	97	59
경남권	158	31	37	39	31	19
경북권	105	21	25	26	21	12
전남권	67	13	15	17	13	8
전북권	37	7	9	9	7	4
충남권	73	14	17	18	14	9
충북권	31	6	7	8	6	4
강원권	30	6	7	8	6	4
제주권	11	2	3	3	2	1
	1,000	199	235	248	198	120

<표 5>를 살펴보면 수도권의 녹음 인원이 전체 수집 인원의 50%에 가깝다. 그러나 수도권의 경우 기구축 DB 및 공개 코퍼스에 상대적으로 많은 음성 자료가 있기 때문에 이 연구에서는 수도권의 인원을 30%로 줄여 그만큼의 비율을 다른 방언권의 조사에 할당한다. 또한 충북권, 강원권, 제주권의 경우 비율로 환산하였을 때 매우 적은 수의 자료만을 수집하게 된다. 따라서 음성 특성을 추출해 낼 수 있는 최소한의 인원을 확보하기 위해 각 방언 및 연령에 따른 피험자의 수를 최소 10명 이상으로 조정하였다. 이와 같이 수도권의 비중을 줄이고, 지역별 최소 인원 확보를 위해 조정한 인구 통계 수치를 <표 6>에 제시하였다.

표 6. 지역별 최소 인원 확보를 위한 조정 인구 통계(1,000명당)

	1,000명당	20대	30대	40대	50대	60대
수도권	300	60	70	70	60	40
경남권	160	30	40	40	30	20
경북권	140	30	35	35	25	15
전남권	100	20	25	25	20	10
전북권	50	10	10	10	10	10
충남권	100	20	25	25	20	10
충북권	50	10	10	10	10	10
강원권	50	10	10	10	10	10
제주권	50	10	10	10	10	10
	1,000	200	235	235	195	135

조사 대상자는 '해당 지역에서 태어나 해당 지역에서 계속 살아왔으며, 현재 해당 지역에 거주하고 있는 사람'이다. 이때 해당 방언권이 아닌 다른 지역에 거주한 시기가 3년 이내인 경우는 피험자로 선정할 수 있도록 한다. 단, 초·중·고등학교 시기에 다른 방언권 거주 경험이 있는 경우는 제외한다. 해당 지역에서 계속 살아왔다는 것은 그 지역에서 나고 자랐다는 뜻

으로서, 초·중·고 학령기를 그곳에서 지낸 사람을 말한다고 볼 수 있다. 학령기는 밀도 높은 교육을 받는 시기로서 학교와 주변 사람들로부터 언어적 영향을 많이 받으므로 해당 시기에 타 방언권에서 거주한 경험이 있는 경우 피험자로 삼지 않은 것이다. 또한 남성의 경우 군복무로 인해 2~3년 타지 생활을 하는 것이 일반적이고, 어학연수 등으로 타지 거주 경험이 있는 20-30대 화자가 많으므로 3년 이내의 타 지역 거주 경험을 허용하여 피험자를 선정한다.

3.2 자료 수집

이 연구는 간접 수집과 직접 수집의 두 가지 방법으로 한국인 표준 음성을 수집하였다. 간접 수집은 공개 코퍼스 또는 고려대학교 민족문화연구원 음성언어센터에서 그간 구축해 놓은 음성 DB에서 음원을 수집하여 이 연구에서 활용할 수 있도록 가공하는 것이고, 직접 수집은 이 연구에서 개발한 프로토콜에 따라 피험자 1인당 10분 이상의 음성 자료를 수집하는 것이다.

3.2.1 간접 수집

공개 코퍼스는 연구 목적으로 활용할 수 있도록 공개되어 있는 코퍼스를 말한다. 주로 국립국어원에서 구축한 것이며, 그 외 기관에서도 활용 가능한 코퍼스가 일부 구축되어 있다. <표 7>에 제시한 바와 같이 공개 코퍼스에서 서울 방언 남녀 화자 140명분의 자료와 방언 음성 자료를 수집할 수 있다.

표 7. 공개 코퍼스 개관

	서울말 낭독체 발화 말뭉치	외국인의 한국어 발화 음성 DB
기관	국립국어원	SiTEC
내용	소설, 수필, 논설 등 930 종류의 문장	단문 10개, 대화문 20개, 단어 88개
인원	20대~60대 남녀 화자 총 120명	한국인 남녀 화자 20명

국립국어원에서 개발한 '서울말 낭독체 발화 말뭉치'는 DVD의 형태로 무료 배포되어 있다. 이 음성 코퍼스는 20대 남녀 화자 각 20명, 30대 남성 화자 20명, 40대 여성 화자 20명, 50대 이상 남녀 화자 각 20명으로 구성된 총 120명의 화자를 대상으로 구축된 것이다. 이 가운데 20~40대 화자 80명은 930 종류의 문장을, 50대 이상 화자 40명은 404개의 문장을 낭독하였다. 실제 구축된 코퍼스에는 30대 남성 화자 1명과 50대 여성 화자 1명의 자료가 유실되어 있어 결과적으로 총 118명분의 자료를 활용할 수 있다.[12)]

이 코퍼스는 음성과 대본이 잘 갖추어져 있고, 녹음 연도, 녹음지, 녹음 환경 및 장비 등에 대한 자료도 함께 기록되어 있으며, 피험자 정보도 비교적 명확하게 제시되어 있어 본 연구에서 활용할 수 있었다.

'외국인의 한국어 발화 음성 DB'는 산업자원부 출연 원광대학교 음성정보기술산업지원센터(SiTEC)에서 산업기술기반 조성 사업의 일환으로 "표준화된 음성 DB의 구축 및 보급"을 위하여 제작한 것이다. 이 가운데 외국인 화자와의 대조를 위해 함께 수집한 한국인 남녀 화자 20명에 대한 자료를 이 연구에서 활용할 수 있다. 이 연구의 화자는 20-30대 서울말 화자로, 각 화자들은 단문 10개, 대화문 20개, 단어 88개 등을 낭독하였다

기구축 DB는 고려대학교 민족문화연구원 음성언어센터에서 구축하여 보유하고 있는 음성 DB를 말한다. <표 8>은 음성언어센터의 음성 DB 중 이 연구에 활용할 수 있는 DB를 제시한 것이다. 화자 식별 시스템 개발의 데이터로 사용되기 위하여 한 화자의 음성을 최소한 3분 이상 확보할 수 있어야 하므로, 음성언어센터의 DB 중에서 '말글비교실험 DB'와 '감정 DB'가 이러한 조건을 가장 잘 충족한다. 앞서 소개한 자유 발화 자료의 경우, 3인이 한 자리에서 자유롭게 대화한 자료이기 때문에, 각 화자의 음성을 분리해 내는 것이 어렵다.

표 8. 음성언어센터 보유 코퍼스 개관

말글비교실험 DB	
내용	2가지 종류의 말하기 과업 수행
인원	20·30대 남녀 화자 총 130명(2008년) 20·30대 남녀 화자 총 74명(2010년)
규모	약 1,200분 분량(2008년) 약 600분 분량(2010년)
감정 DB	
내용	평서문 36개, 의문문 14개, 명령/청유문 15개 총 65개 문장을 4가지 감정으로 2회씩 발화(2005년) 평서문, 의문문, 명령/청유문, 기타 각 10개씩 총 40개 문장을 4가지 감정으로 2회씩 발화(2006년)
인원	20대 남녀 화자 각 5명, 총 10명(2005년) 20대 남녀 화자 각 10명, 총 20명(2006년)
규모	총 5,200문장(2005년) 총 6,400문장(2006년)

'말글비교실험 DB'는 동일 화자가 동일 주제로 수행한 말하기와 글쓰기 활동의 결과물을 수집하여 구어와 문어의 특성을 고찰할 수 있도록 설계한 DB이다. 이 DB에서 20·30대 서울 방언 남녀 화자의 음성 총 204명분을 얻을 수 있었다. 각 화자는

12) 118명 가운데 18명의 자료(20대 남성 4명, 20대 여성 1명, 30대 남성 2명, 40대 여성 4명, 50대 이상 남성 1명, 50대 이상 여성 6명)는 낭독 문장 중 일부가 유실되었다. 대체로 1-2개의 문장이 유실되었으나, 20대 남성 화자 2명의 경우 930 문장 가운데 각각 879개, 792개의 문장이 유실되었다.

3~5분의 발화를 2번 수행하였고, 발화 수행 시 적용하도록 유도한 사용역은 면접 상황, 방송 보도 상황, 수업 발표 상황 등으로 다양하다.

‘감정 DB’는 평상, 기쁨, 화남, 슬픔 등 4가지 감정을 표현할 수 있는 대화 형식의 문장을 발화한 자료이다. 해당 DB에서 서울 방언 20대 남녀 화자 총 30명분의 자료를 활용할 수 있다. 각 화자는 다양한 종결법의 문장을 4가지 감정으로 2회씩 발화하였다.

이렇게 하여 공개 코퍼스와 본 센터의 기구축 DB를 통한 간접 수집으로 총 373명분의 음성을 수집하였다. <표 9>는 간접 수집 자료의 수를 성별과 연령에 따라 나타낸 것이다. 기존 대부분의 코퍼스가 그러하듯, 모든 자료가 서울말 화자를 대상으로 수집한 것이기 때문에 지역적으로는 수도권에만 한정되어 있다. 또한 20대가 전체 인원의 73.2%에 해당할 정도로 연령의 편중이 심했다.

표 9. 간접 수집 자료의 규모(단위: 명)

	남	여	합계	%
20대	129	144	273	73.2
30대	33	5	38	10.2
40대	2	22	24	6.4
50대	11	16	27	7.2
60대 이상	9	2	11	2.9
합계	184	189	373	100.0

3.2.2 직접 수집

직접 수집의 진행 과정은 크게 녹음 전, 녹음, 녹음 후 과정의 세 단계로 나뉜다. 먼저 녹음 전에는 조용한 공간을 확보하고, 조건에 맞는 피험자를 섭외하여 서류를 작성하게 함으로써 피험자 관련 정보를 확보한다. 피험자들이 작성하는 서류는 ‘피험자 동의서’와 ‘피험자 정보’이다. ‘피험자 서면 동의 설명문’을 제시한 후 충분히 읽게 하여 동의 여부를 작성하게 한다. 동의 설명문에 동의한 경우에만 피험자 정보 수집을 위한 서류를 작성하게 하였다. 녹음 전 과정이 끝나면 녹음 과정에 들어가게 된다. 녹음은 5가지 발화 과제를 차례로 수행하며 이루어진다. 녹음 후에는 피험자에게 소정의 사례비를 지급하고 ‘녹음비 수령 영수증’을 받으며, ‘녹음비 수령자 목록’을 작성한다. 전체 과정을 거치는 데 소요되는 시간은 피험자 1인당 약 30분이었다.

1) 조사원

지역별 직접 수집을 위해서는 각 지역에서의 수집 활동이 필요하다. 이 연구에서는 각 지역 대학의 국어 관련 전공 학과에 재학 중인 대학생 또는 대학원생을 대상으로 조사원을 모집하였다. 모집된 조사원들에게는 조사원 교육을 실시하여 피험자의 요건과 녹음 장비 사용법, 자료 수집 절차 등을 숙지하도록 하였다. 20-30대 학생 조사원들은 또래 화자들을 많이 조사할 수 있고 음성 파일 처리나 전사 등 후처리를 쉽게 한다는 장점이 있지만, 장년층 화자의 섭외에 어려움이 있을 수 있다. 그런 이유로 연령대별 균형이 맞지 않은 일부 지역에서는 장년층 조사원을 추가로 섭외하여 학생 조사원들이 충분히 수집하지 못한 장년층 자료를 수집하였다. 장년층 조사원들은 학생 조사원들에 비하여 높은 연령층과의 접촉이 용이하기 때문에 40대 이상의 피험자 섭외에 어려움을 덜 겪는 것으로 드러났다. 다만 학생들은 방학 기간을 이용해 집약적으로 조사 활동을 할 수 있으나, 장년층 조사원은 직업이 있는 경우 조사 활동을 할 수 있는 시간적 여유가 부족할 수 있다. 따라서 조사원 모집에 있어서 크게 두 연령층으로 나누어 모집하고, 시기를 나누어 조사 활동을 하도록 하는 방안이 효과적이다.

2) 수집 자료

직접 수집에서는 피험자 1인당 최소 10분 이상의 발화를 수집한다. 음성 언어 수집에는 주어진 대본을 활용한 낭독 발화와 자연스러운 발화를 녹음하는 자유 발화 두 가지 유형이 있다. 낭독 발화는 대본을 제시하고 그대로 읽도록 하는 것이므로 모든 피험자로부터 동일한 내용의 음성을 얻기에 유용한 방법이다. 한편 자유 발화는, 주어진 대본을 전형적인 운율로 읽게 되는 낭독 발화와는 달리, 최대한 자연스럽고 자발적인 발화의 음성을 수집하는 데에 그 목적이 있다. 이 연구에서는 이와 같은 각 발화 유형의 장점을 모두 포괄하기 위하여 낭독 발화와 (준)자유 발화로 <그림 1>과 같이 총 5가지 발화 과제를 구성하였다.

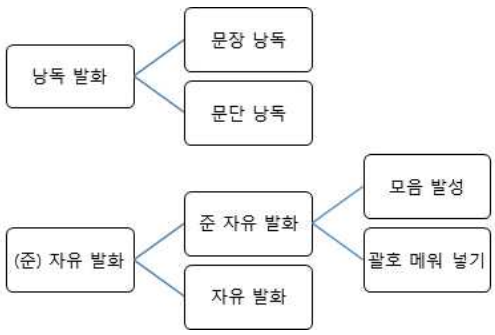

그림 1. 직접 수집 발화 자료 구성

낭독 발화의 대본은 문장과 문단의 두 가지로 마련된다. 문장은 한국어의 모든 음운이 고루 실현되도록 개발한 55개 문장이다. 55개 문장은 한국어 분절음(자음 19개, 단모음과 이중모음 21개)을 수집하기 위한 문장 40개와, 특징적인 음운 현상을 관찰하기 위한 문장 15개로 구성하였다. 한국어 분절음을 조사하기 위한 문장의 경우, 목표 분절음 40개가 문장의 첫음절에 나타나며, 한 문장에 목표 분절음이 최소 3번씩 실현되도

록 구성하였다. 음운 현상의 경우, 겹받침의 발음, ./ㄴㄹ/ 연쇄의 발음, 경음화, /ㄴ/ 삽입 등과 같이 방언, 연령, 성별에 따라 차이가 있는 것으로 선행 연구에서 논의되었던 음운 현상을 관찰할 수 있는 단어로 구성하였다. 문장 낭독의 경우, 목표 분절음을 발음한 음성을 수집해야 한다는 목적이 있으므로 피험자가 잘못 발음한 경우 다시 발음하도록 하였다. 다만 낭독 중간에 그러한 개입이 있을 경우 피험자의 집중력을 해치고, 심리적 부담감을 가중시킬 수 있으므로 전체 문장의 낭독이 모두 끝난 후, 다시 읽어야 할 문장의 번호를 일러 주어 틀린 문장만 다시 읽도록 하였다. 본 연구를 위해 설계한 단모음 조사용 문장과 자음 조사용 문장의 일부를 보이면 다음과 같다. 전체 실험 문장은 부록에 제시하였다.

ㅏ: 아픈 강아지를 안고 아버지와 병원에 갔더니 늑막염이라고 했다.
ㅐ: 애타는 마음으로 백 일 동안 매일 너를 기다렸다.
ㅓ: 어머니는 언제 돈을 벌어 거기서 돌아오실까?
ㅔ: 에누리 없이 파는 가게에서 어제 제비를 뽑았다.
ㅜ: 우울한 날에는 구름을 보며 우유를 마신다.
ㄱ: 가평의 군부대에서 각자 군 생활을 했다.
ㄴ: 난민들은 누더기 옷을 입고 나무 아래서 노닌다.
ㄷ: 두더지는 다시 찬란한 땅을 디딜 수 없었다.
ㄹ: 리본에 루비를 달고 솜이불 위에서 라면을 먹었다.

낭독 음성은 문단 단위에서도 수집하였다. 문단 낭독 음성에서는 문장보다 긴 단위에서 나타나는 운율적 특성을 관찰할 수 있게 한다. 본 연구에서는 운율적 특징은 물론, 음성학적, 음운론적 특징을 관찰하기 용이하도록 설계된 3개의 문단을 개발하였다. 이 가운데 2개의 문단은 공명음의 비율을 높이고 음운구 단위의 음절수를 대체로 3~4음절로 구성하여 발화 전체적으로 전형적인 한국어 운율이 자연스럽게 드러날 수 있도록 하였다. 다른 하나의 문단은 일상적이고 친숙한 내용으로 쉽게 낭독할 수 있도록 설계하였다. 문단 낭독 과제에서는 다소 잘못 읽은 부분이 있더라도 자연스러운 운율로 낭독한 경우에는 다시 읽게 하지 않았다. 다만, 문단의 일부를 누락하여 발화하거나, 자연스러운 발화라도 제시문과 다른 부분이 지나치게 많은 경우 해당 문단을 다시 읽도록 요청하였다. 공명음 위주로 구성된 문단 중 하나를 보이면 다음과 같다.

남일이네 야옹이는 멍멍이를 미워합니다. 야옹이는 멍멍이의 마음을 모릅니다. 그래서 멍멍이랑 놀아주지 않습니다. 은행나무 위에는 야옹이만 올라옵니다. 무모한 멍멍이는 나무 위로 날아오릅니다. 그렇지만 너무 높아서 오르기가 어렵습니다. 야옹이는 매일매일 나무 위에 머무릅니다. 위에서 얄미운 울음만 웁니다. 나무 아래 누워있는 멍멍이는 무료합니다. 야옹이는 야밤에만 아래로 내려옵니다. 우울한 멍멍이는 애먼 나를 원망합니다.

(준)자유 발화는 준자유 발화 2가지와 자유 발화로 구성하였다. 준자유 발화는 모음 발성과 괄호 메워 넣기의 두 가지 과제로 구성하였는데, 대본을 낭독하는 것은 아니지만 주어진 조건이나 자료를 바탕으로 하여 발화를 한다는 점에서 준자유 발화라고 할 수 있다.

먼저 모음 발성은 각 피험자의 음성을 음성학적 측면에서 정밀하게 측정하기 위하여 포함하였다. 단모음 /ㅏ, ㅐ, ㅓ, ㅔ, ㅗ, ㅜ, ㅡ, ㅣ/ 8개를 자신에게 가장 편안한 음높이와 크기로 3초씩의 연장 발성을 3회 반복하도록 하였다. 모음 연장 발성의 수집은 음성치료 분야에서 음성장애 진단 및 평가에 활용될 수 있을 것으로 기대한다. 음성의 질에 대한 음향적 평가를 위해서는, 낭독이나 대화 등 지속적인 발화 음성에 비해 모음만을 발성한 음성이 유용하기 때문이다[7].

다음으로 괄호 메워 넣기 방식의 대본을 활용하여 준자유 발화 형식의 음성을 수집하였다. 괄호 메워 넣기 대본은 “제 이름은 (　　)입니다.”와 같이 빈칸이 있는 문장들로 구성되어 있어서, 피험자들이 세부사항을 직접 채워 넣으며 말하도록 설계되었다. 이러한 방식의 장점은 두 가지가 있다. 첫째, 비교적 쉽게 자발적 발화를 수집할 수 있다는 점이다. 혼자서 아무런 대본 없이 일정 시간 이상 발화하도록 하면, 대부분의 피험자들은 발화의 주제 선택에서부터 심리적 부담을 느껴서 일정 시간 이상 발화를 지속하는 데 어려움을 보인다. 괄호 메워 넣기 방식은 비록 완전한 자유 발화는 아니지만 자발적인 발화에 가까운 발화를 짧은 시간 안에 원하는 분량만큼 얻을 수 있다. 둘째, 모든 피험자가 비슷한 주제와 내용에 대해 자발적 발화를 하도록 할 수 있다는 점이다. 따라서 제각기 다른 주제를 이야기한 자유 발화 자료에 비해 화자 간에 보다 정밀한 비교를 시도할 수 있는 자료가 된다. 이 연구에서는 ‘신상, 가족/친구, 지역/교통, 여가/문화, 상식’의 5개 주제를 가지고 한 주제당 4~6문장을 구성하여 괄호 메워 넣기 발화를 유도하였다. 이 가운데 ‘여가/문화’ 부분의 실험 문장을 예시로 보이면 다음과 같다.

여가/문화
▸제 취미는 (_____)입니다.
/ 저는 (_____) 하는 것을 좋아합니다.
▸제가 좋아하는 스포츠는 (_____) 입니다.
▸운동선수 중에는 (_____)를 좋아합니다.
/ (딱히 좋아하는 사람이 없습니다).
▸제가 제일 좋아하는 음식은 (_____)입니다.
▸중국집에서는 (짜장면/짬뽕)을 먹고,
치킨은 (양념/프라이드)를 먹습니다.
▸가장 최근에 갔다 온 여행은
_____(어느 나라) ____(어느 도시)로 갔던 여행입니다.
▸(__)박 (__)일 동안 (혼자서/___와) 여행했었고,
숙소는 (호텔/콘도/민박/...)이었습니다.

자유 발화는 독백 형태의 자발적 발화로서, 피험자 1인당 3~5분 정도 자유롭게 이야기하게 하여 녹음한 자료이다. 기본적으로는 피험자가 주제를 선택하여 자유롭게 발화하도록 하였다. 다만, 피험자가 주제 선택을 어려워하는 경우, 가족 소개, 아끼는 물건, 지난 주말에 한 일, 좋아하는 영화의 줄거리 등 쉽게 이야기할 수 있는 주제를 제시한다. 제시한 주제로도 이야기하는 것을 어려워하는 경우 그림 자료(전래 동화의 줄거리를 표현하는 4단 그림, 가족사진 등)를 제시하여, 그림을 보고 설명하는 과제를 수행하도록 하였다.

한편, 다른 과제와는 달리 자유 발화는 음성 자료에 대응되는 텍스트 자료가 사전에 준비되지 않으므로, 사후에 전사를 해야 한다. 따라서 이 연구의 자유 발화 수집에는 철자 전사 절차가 포함된다. 전사 후 그 분량이 A4용지의 절반 이상이 되도록 하며, 그 분량은 대체로 250어절 이상이었다. 자유 발화 녹음 시에는 전사 후 250어절 이상의 발화 분량을 확보할 수 있도록 발화 중간에 들어간 쉼이나 발화 속도 등을 고려하여 녹음 분량을 조절해야 한다.

3) 수집 과정

과제의 제시 순서는 <그림 2>와 같이 모음 발성, 문단 낭독, 문장 낭독, 괄호 메워 넣기, 자유 발화 순이었다. 이와 같은 순서는 과제 수행에 대한 피험자의 심리적 부담이 낮은 것에서부터 시작하여 점차 피험자의 능동적 참여가 필요한 것으로 진행하게 함으로써 피험자가 과제를 쉽게 수행할 수 있도록 설계된 것이다.

그림 2. 발화 과제 수행 순서

20대에서 60대 이상의 피험자들을 대상으로 하면서도 일관된 수집을 해야 하므로 해당되는 모든 연령대의 피험자들이 수행할 수 있는 발화 과제가 될 수 있도록 설계하였다. 하지만 연령대에 따른 과제 수행 능력 차이를 고려하여 40대 이하와 50대 이상의 피험자에게 부여하는 과제에 약간의 차이를 두었다. 예비 실험 결과 50대 이상의 경우 낭독 과제에서 어려움을 느끼고 소요 시간 또한 긴 것이 확인되었다. 따라서 낭독 과제의 대본을 보기 쉽게 편집하고, 대본의 분량을 줄였다. 하지만 문장 낭독 과제의 경우 모든 문장이 특정 분절음이나 음운 현상을 확인하려는 목적을 가지고 있으므로 모두 낭독하게 하는 것이 바람직하다고 판단하였다. 따라서 문장 과제가 아닌 문단 과제에서 과제의 양을 줄여, 50대 이상의 피험자는 3개의 문단이 아닌 1개의 문단을 낭독하도록 하였다. 그리고 낭독 발화와 자유 발화의 중간 성격을 가지는 괄호 메워 넣기에서도 5개 문단에서 3개 문단으로 분량을 줄임으로써 부담을 경감하였다. 따라서 50대 이상의 피험자는 문단 낭독과 괄호 메워 넣기에서 다른 연령대에 비해 적은 양의 과제를 수행하였다.

음성 왜곡이 나타나거나 잡음 등이 포함되지 않도록 하기 위해 고성능 디지털 녹음 장비를 활용하여 녹음을 수행하였다. 녹음에 사용한 녹음기는 TASCAM DR-07과 SONY PCM-D50, SONY PCM-M10 3종이며, 삼각대를 사용하여 녹음기를 고정한 후 녹음기의 내장 마이크를 사용하여 녹음하였다 44.1kHz 표본추출률, 16bit 양자화로 녹음하고 wav 형식(format) 파일로 저장하였다 음성 자료는 녹음실 혹은 조용한 녹음 공간을 확보하여 잡음을 최소화한 환경에서 수집하였다.

이 연구는 총 1,000명의 음성을 수집하는 것을 목표로 하였다 간접 수집에서 373명의 음성을 수집할 수 있으므로 직접 수집에서는 최소 627명의 음성을 수집해야 한다. 연구 결과 총 639명의 자료를 직접 수집을 통해 확보할 수 있었다. 직접 수집 결과를 지역 및 연령에 따른 인원수로 나타내면 <표 10>과 같다. 직접 수집 639명과 간접 수집 373명을 합한 총 1,012명의 수집 결과는 <표 11>에 정리하였다.

수집 결과를 목표 인원에 대비해 보면 지역별로는 수도권과 전북은 초과 달성, 나머지 지역은 미달로 요약할 수 있다. 연령별로는 20대는 초과 달성하였으나, 높은 연령으로 갈수록 조사가 미진하게 이루어진 것을 확인할 수 있다. 성별에 따라서는 남성과 여성이 비교적 고르게 수집되었다. 이후의 연구에서는 이 결과를 고려하여 조사가 미진한 지역 및 고연령대의 자료 수집에 역점을 둘 필요가 있다. 이 연구는 현재 1차 연도를 마

표 10. 직접 수집 자료의 규모(단위: 명)

	20대		30대		40대		50대		60대		합계
	남	여	남	여	남	여	남	여	남	여	
수도권	3	3	19	29	7	15	9	19	2	7	113
경남권	20	29	8	10	4	15	14	17	2	2	121
경북권	15	27	2	7	13	24	2	9	3	3	105
전남권	16	19	0	2	1	27	1	9	0	0	75
전북권	13	16	8	9	5	15	11	3	2	2	84
충남권	15	15	5	12	3	3	1	0	6	4	64
충북권	1	6	0	0	1	1	1	2	0	0	12
강원권	3	10	2	1	2	3	4	3	1	1	30
제주권	2	13	12	2	1	0	4	1	0	0	35
합계	88	138	56	72	37	103	47	63	16	19	639

표 11. 직접 수집과 간접 수집을 합한 전체 조사 자료의 규모(단위: 명)

	20대		30대		40대		50대		60대		합계
	남	여	남	여	남	여	남	여	남	여	
수도권	132	147	52	34	9	37	20	35	11	9	486
경남권	20	29	8	10	4	15	14	17	2	2	121
경북권	15	27	2	7	13	24	2	9	3	3	105
전남권	16	19	0	2	1	27	1	9	0	0	75
전북권	13	16	8	9	5	15	11	3	2	2	84
충남권	15	15	5	12	3	3	1	0	6	4	64
충북권	1	6	0	0	1	1	1	2	0	0	12
강원권	3	10	2	1	2	3	4	3	1	1	30
제주권	2	13	12	2	1	0	4	1	0	0	35
합계	217	282	89	77	39	125	58	79	25	21	1,012

친 것이고 향후 2차와 3차 연도에도 연구가 지속될 예정이므로, 수집 현황에 따라 수집 계획을 조정함으로써 최종 말뭉치에서는 이러한 불균형이 해소될 수 있을 것이다.

4. 연구 결과 활용 방안

이 연구에서 소개한 연구 방법과 내용으로 구축되는 한국인 표준 음성 DB는 인구 통계학적 균형을 갖춘 코퍼스라는 점에서 기존의 한국어 음성 코퍼스와 차별화되는 강점을 가진다. 이 DB는 용의자 음성식별을 위한 한국인 음성 데이터베이스 수집을 목적으로 기획된 것으로, 법과학적 목적으로 유용하게 사용될 것이다. 현재 해당 자료는 용의자 음성식별을 위한 음성 자동분석 시스템 개발의 입력 데이터로 사용되고 있다. 이후 용의자 음성을 분석하여 신속한 화자 식별 및 신원 추정이 가능하게 하기 위한 한국인의 표준 음성 표본으로서의 역할을 충실히 수행할 것이며, 다양한 범죄 수사 및 이와 관련된 법과학적 응용에 적용될 것이다.

또한 한국인 표준 음성 DB는 범용 DB로서, 언어학의 다른 분야 및 기타 응용 분야에서도 활용될 것이 기대된다. 지역에 따른 음성을 비교하는 방언학은 물론, 성별·지역·세대 등에 따른 언어 차이를 연구하는 사회 언어학 분야에도 연구 자료로서 훌륭한 역할을 할 것이다. 또한 언어 병리학 분야에서의 표준 데이터로서도 활용될 수 있다. 나아가 기초 연구들을 통하여 표준 발음 정책 등 언어 정책의 근거 자료로도 사용될 수 있다.

한편 연구 과정에서 음성 DB를 구축하기 위하여 체계적인 음성 수집 프로토콜이 개발되었으며, 이 프로토콜 또한 이후 연구에 활용될 것이 기대된다. DB 구축 초기에는 DB 설계와 프로토콜 개발에 상당한 노력이 들어가는 만큼, 이 연구를 통해 개발된 음성 수집 프로토콜과 수집 과정의 다양한 고려점들은 추후 음성 DB 구축 시 활용될 수 있을 것이다.

또한 이 연구에서 개발한 녹음용 대본 중 괄호 메워 넣기 대본은 자발적 발화 자료를 필요로 하는 대부분의 음성 연구에 활용될 수 있다. 분절음별 발음을 관찰하기 용이하게 개발된 문단과 문장을 통해 한국어의 분절음적 측면은 물론 운율적 측면과 음운 현상 등을 관찰할 수 있어서 음성학과 음운론 분야에서 유용하게 활용될 것이다. 또한 언어 병리학, 음성 공학 등 주요 응용 분야에서 검사 및 평가 기준으로 사용될 수 있다. 특히, 본 연구의 수집 자료 중에는 음향적 평가를 위한 모음 연장 발성 자료도 포함되어 있어 음성치료 분야의 연구 및 활용에도 기여할 수 있다. 그뿐만 아니라 화자의 다양성과 대표성을 고려하고 동일 화자의 음성을 여러 양식으로 수집하였다는 점에서, 화자 인식 시스템 개발 분야에서도 유용하게 사용될 수 있다.

5. 결론

이 연구는 한국인 표준 음성 DB의 구축에 대한 구체적인 설계 방안과 자료 수집 과정에서의 고려점 등을 구체적으로 논의하였다. 인구 통계학적 비례에 따라 표본 수집 대상 모집단을 선정한 이 음성 코퍼스는 한국인 표준 음성 DB라 할 만한 대표성을 가질 것이다. 또한 낭독 발화와 자유 발화를 고루 수집하여 자료 유형의 균형성 또한 확보할 것이다. 이 연구는 또한 각 음성 수집의 목적에 부합되는 실험 자료를 개발하였다. 그리고 실제 음성 수집 과정에 필요한 프로토콜을 개발하여, 전국에 걸친 대규모 조사임에도 일관성이 유지되도록 하였다.

또한 이 연구에서는 한국인의 표준 음성 수집을 위하여 낭독 발화와 (준)자유 발화로 구성된 발화 과제를 개발하였다. 낭독 발화는 다시 문장 낭독과 문단 낭독으로 세분되며, (준)자유 발화는 모음 발성, 괄호 메워 넣기, 자유 발화로 세분된다. 이러한 실험 자료 구성은 다양한 발화 형식을 포함한 풍부하고 균형 있는 음성 수집을 가능하게 하여, 이 연구에서 구축되는 음성 코퍼스의 활용도를 높일 수 있을 것으로 기대한다.

이와 같은 방법을 통하여 구축된 한국인 표준 음성 DB는 기존의 어떤 한국어 음성 코퍼스와도 다른 강점을 지닐 것이다. 그러므로 이 DB는 음성학 등 언어학 연구뿐 아니라 주요 응용 분야에서 다양한 목적으로 사용될 수 있을 것으로 기대된다. 특히 5단계로 구성된 실험 자료는 한국어 음성을 다각적으로 분석할 수 있는 음성 코퍼스의 구축을 가능하게 함으로써 여러 분야에서 요구되는 대규모 음성 코퍼스로서 그 기능을 충실히 수행할 수 있을 것이다. 이를 위해 이후의 연구에서도 연구 방법과 내용을 지속적으로 보완하여 DB 규모를 확대할 계획이다.

참고문헌

[1] 서상규·김형정 (2005). 구어 말뭉치 설계의 몇 가지 조건, 언어사실과 관점, 14, 5-29.

[2] 국립국어원 (2007). 21세기 세종계획 국어 특수자료 구축, 서울: 국립국어원.

[3] 윤원희 외 (2013). 한국어 자연발화 음성코퍼스 구축을 위한 기초 연구, 실험음성학연구회 강독회.

[4] Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability, *Speech Communication*, 45(1), 89-95.

[5] Nolan, F., de Jong, G. and McDougall, K. (2006). Introducing the DyViS project: Dynamic variability in speech: a forensic phonetic study of British English, In *Abstract Proc. Annual Conf. of the International Association for Forensic Phonetics and Acoustics(IAFPA)*.

[6] Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond, *Speech Communication*, 9(4), 351-356.

[7] Parsa, V., & Jamieson, D. G. (2001). Acoustic Discrimination of Pathological Voice: Sustained Vowels Versus Continuous Speech, *Journal of Speech, Language, and Hearing Research*, 44(2), 327-339.

• **신지영 (Shin, Jiyoung)**
고려대학교 국어국문학과
서울시 성북구 안암로 145
Tel: 02-3290-1973
Email: shinjy@korea.ac.kr
관심분야: 음성학, 음운론

• **장혜진 (Jang, Hyejin)** 교신저자
고려대학교 국제한국언어문화연구소
서울시 성북구 안암로 145
Tel: 02-3290-2908
Email: jina49@korea.ac.kr
관심분야: 음성학, 음운론, 방언학

• **강연민 (Kang, Younmin)**
고려대학교 민족문화연구원 음성언어센터
서울시 성북구 안암로 145
Tel: 02-3290-2505
Email: flour@korea.ac.kr
관심분야: 음성학, 음운론
현재 국어국문학과 대학원 박사과정 재학 중

• **김경화 (Kim, Kyung-Wha)**
대검찰청 과학수사담당관실
서울시 서초구 서초동 반포대로 157
Tel: 02-3480-2150
Email: savoix@spo.go.kr
관심분야: 법음성학, 화자 인식

부록

1. 문단 대본

미영이랑 나연이는 단짝입니다. 미영이와 나연이는 노래하며 놉니다. 마루 위에 나란히 누워 낭랑히 노래합니다. 나연이는 노래를 매우 많이 압니다. 노래도 더 잘해서 미영이에게 알려 줍니다. 미영이는 음악에 어울리는 안무를 마련합니다. 어느 날 나연이는 미영이를 놀립니다. 자기보다 노래를 못한다고 놀립니다. 미영이는 남몰래 노래를 연마합니다. 미영이의 능력이 나날이 늘어납니다. 그래서 나연이는 더 이상 미영이를 놀릴 수 없게 되었습니다. 나연이는 사과했고 둘은 다시 사이가 좋아졌습니다.

나는 라면을 매우 좋아한다. 생라면도 컵라면도 좋지만 뭐니 뭐니 해도 끓인 라면이 제일 좋다. 양은냄비를 꺼내 가스레인지 위에 올리고, 물이 끓을 때까지 조리법을 읽는다. 물이 보글보글 끓기 시작하면 면과 스프를 넣고 끓인다. 계란 노른자가 익는 모습을 보고 있을 때가 가장 즐거운 순간이다. 얼른 먹고 싶어서 군침을 꿀떡꿀떡 삼키면서도, 조리 시간을 지키는 것이 나의 철칙이다. 열과 성을 다해 만든 라면을 한 젓가락 먹으면 절로 미소가 난다. 입에서 김이 호호 나오고 땀이 삘삘 나지만 젓가락질을 멈출 수가 없다. 그야말로 무아지경에 빠지고 마는 것이다.

남일이네 야옹이는 멍멍이를 미워합니다. 야옹이는 멍멍이의 마음을 모릅니다. 그래서 멍멍이랑 놀아주지 않습니다. 은행나무 위에는 야옹이만 올라옵니다. 무모한 멍멍이는 나무 위로 날아오릅니다. 그렇지만 너무 높아서 오르기가 어렵습니다. 야옹이는 매일매일 나무 위에 머무릅니다. 위에서 얄미운 울음만 웁니다. 나무 아래 누워있는 멍멍이는 무료합니다. 야옹이는 야밤에만 아래로 내려옵니다. 우울한 멍멍이는 애먼 나를 원망합니다.

2. 문장 대본

1 사자가 수풀 속에 실체를 숨겼다.
2 윗집의 귀여운 동생을 위해서 귀찮지만 옷 입고 뒷마당으로 나갔다.
3 하늘이는 후미진 골목에서 희미하게 웃으며 히죽거렸다.
4 일요일날에는 안암 1동에서 2동으로 이동하는 것도 은근히 일이다.
5 얘는 얘기꾼이라 얘기를 잘 해.
6 소주와 김밥과 통닭을 즐겨 먹은 까닭에 수일 내에 돼지가 될 것 같다.
7 이쪽에서 우는 아기의 이름을 정하자.
8 예시로 문예 창작반의 월례 행사인 섬유 공예를 체험해 보자.
9 서른여덟의 김유신은 권력을 이용해 불법으로 생산라인을 개조하였다.
10 넓게 지어진 연륙교가 효과적으로 제 몫을 다 해내고 있다.
11 뚜껑이 검은색을 띠어서 따로 설거지했다.
12 의사인 나의 말을 의심하지 마십시오.
13 서울역에서 본 손예진의 첫인상은 넋이 나갈 정도였다.
14 차가운 추어탕과 떫은 치약, 고추를 썰어 만든 과자를 샀다.
15 부드러운 바닷바람을 맞으며 밥집에 들어갔다.
16 큰 빚을 져서 찜닭도 못 먹고 땀이 맺히도록 밤낮으로 뛰어다녔다.
17 애타는 마음으로 백 일 동안 매일 너를 기다렸다.
18 늦여름이나 가을날에 상견례를 하려고 온라인으로 식당을 예약했다.
19 음운론을 가르치시던 담임 선생님은 흙에서 넓죽한 고등어를 캐셨다.
20 야구 방망이가 없으니 약간 갸름한 것을 사자.
21 웬일인지 웰빙에 대한 궤변을 늘어놓지 않았다.
22 이 모 씨의 이모가 마침내 고소 절차를 밟게 되었다.
23 에누리 없이 파는 가게에서 어제 제비를 뽑았다.
24 난민들은 누더기 옷을 입고 나무 아래서 노닌다.
25 유구한 역사를 가진 유서 깊은 가문의 규수였다.
26 아픈 강아지를 안고 아버지와 병원에 갔더니 늑막염이라고 했다.
27 은사님의 은혜에 그저 울먹이기만 했다.
28 아기 옷을 벗기고 입히면서 빗으로 머리도 빗겨 주었다.
29 완두콩을 과식하다 완전히 체해서 수레를 끌어 병원에 갔다.
30 우울한 날에는 구름을 보며 우유를 마신다.
31 왜 괜찮으니 괘념치 말라고 하는지 이해가 안 된다.
32 밝은 빛으로 곱하기 공부를 하다가 급하게 책 한 권을 읽었다.
33 찜통 같은 쪽방에 쭈그려 앉아 짜증을 냈다.
34 두더지는 다시 찬란한 땅을 디딜 수 없었다.
35 타조는 투명한 유리에서 칼날과 티끌을 발견했다.
36 외삼촌은 금융업에 종사해서 외국에 굉장히 자주 나간다.
37 요즘에는 요가를 배우는 것이 교사들 사이에서 유행이다.
38 가평의 군부대에서 각자 군 생활을 했다.
39 싸움 같은 씨름을 한 뒤 가자미로 죽을 쑤어 먹었다.
40 원래 강원도는 여기보다 훨씬 춥기로 유명하지.
41 피로를 풀기 위해 푸른 파도를 보며 모래를 밟고 놀았다.
42 역시 어려운 일을 견뎌야 여러 가지를 배울 수 있다.
43 키 큰 기자가 쿵쿵거리는 쿠데타 현장을 카메라로 찍었다.
44 어머니는 언제 돈을 벌어 거기서 돌아오실까?
45 빠르게 뛰다가 나무뿌리에 걸려 발목이 삐었다.
46 저 병에 든 약을 마시면 병에 차도가 있을 것이다.
47 원룸에 사니까 절약할 필요가 없어서 방을 밝게 한다.
48 지금 자전거 가게에서 주민 회의가 진행 중이다.
49 산기슭에 있는 장미꽃으로 장식을 하려다가 가시의 끝을 만졌다.
50 까치는 꾸물거리다가 구멍 사이에 끼었다.
51 사기그릇 가게로부터 사기를 당했다는 걸 알고 그들은 사기가 떨어졌다.
52 리본에 루비를 달고 솜이불 위에서 라면을 먹었다.

53 오늘은 노동절을 맞아 보육원 사람들에게 보답을 했다.
54 밤마다 미로를 헤매며 무서운 마음이 들었다.
55 이를 뽑으려고 대기실에서 기다리고 있었다.

3. 괄호 메워 넣기

신상

제 이름은 _____ 입니다.
저는 _____(쥐/소/호랑이/돼지/...) 띠입니다.
제가 태어난 곳은 _____시(군)이고,
주로 산 곳은 _____시(군)입니다.
출생지인 _____시(군)에서는 ___살까지 살았습니다.
아버지는 _____출신이시고, 어머니는 _____출신이십니다.
제 위로는 (형/누나/오빠/언니) ___명이 있고,
아래로 (남동생/여동생) ____명이 있습니다.

가족/친구

우리 가족은 _____, _____, _____, 그리고 저이고,
그래서 모두 ____명입니다.
지금 저랑 같이 살고 있는 사람은 _____, _____, _____입니다.
가장 친한 친구는 (고등학교/직장/...)에서 만난 친구입니다.
친구와 주로 하는 이야기는 _____에 대한 것입니다.

지역/교통

저는 평소에 주로 (지하철/버스/택시/자가용/...)을
타고 다닙니다.
우리 동네는 교통이 (편리합니다/불편합니다).
우리 집에서 서울역에 가려면 _____을 타고 가야 합니다.
우리 집 근처에는 (마트/시장)이 있는데,
(걸어서/버스로/...) ____분 거리에 있습니다.
우리 지역은 _____(사과/대나무/광한루/...)이 유명합니다.

여가/문화

제 취미는 _____입니다.
/ 저는 _____ 하는 것을 좋아합니다.
제가 좋아하는 스포츠는 _____입니다.
운동선수 중에는 (_____를 좋아합니다
/딱히 좋아하는 사람이 없습니다).
제가 제일 좋아하는 음식은 _____입니다.
중국집에서는 (짜장면/짬뽕)을 먹고,
치킨은 (양념/프라이드)를 먹습니다.
가장 최근에 갔다 온 여행은
____(어느 나라) __(어느 도시)로 갔던 여행입니다.
___박 ___일 동안 (혼자서/ __와) 여행했었고,
숙소는 (호텔/콘도/민박/...)이었습니다.

상식

무지개의 일곱 색깔은 ___, ___, ___, ___, ___, ___, ___입니다.
일주일은 __요일, __요일, __요일,
__요일, __요일, __요일, __요일입니다.
설날은 음력 ___월 ___일이고,
크리스마스는 ___월 ___일입니다.
지금 계절은 _____입니다.
어제는 (비가 왔고/해가 났고/흐렸고...),
오늘은 (비가 옵니다/해가 납니다/흐립니다...).
1년은 _____일이고,
1주일은 ____일이고, 하루는 ____시간입니다.