ISSN 2005-8063
2017. 6. 30.
Vol.9 No.2
pp. 9-22

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

http://dx.doi.org/10.13064/KSSS.2017.9.2.009

한국인 표준 음성 DB 구축(II)*

Developing a Korean standard speech DB (II)

신 지 영**· 김 경 화

Shin, Jiyoung · Kim, KyungWha

Abstract

The purpose of this paper is to report the whole process of developing Korean Standard Speech Database (KSS DB). This project is supported by SPO (Supreme Prosecutors' Office) research grant for three years from 2014 to 2016. KSS DB is designed to provide speech data for acoustic-phonetic and phonological studies and speaker recognition system. For the samples to represent the spoken Korean, sociolinguistic factors, such as region (9 regional dialects), age (5 age groups over 20) and gender (male and female) were considered. The goal of the project is to collect over 3,000 male and female speakers of nine regional dialects and five age groups employing direct and indirect methods. Speech samples of 3,191 speakers (2,829 speakers and 362 speakers using direct and indirect methods, respectively) are collected and databased. KSS DB designs to collect read and spontaneous speech samples from each speaker carrying out 5 speech tasks: three (pseudo-)spontaneous speech tasks (producing prolonged simple vowels, 28 blanked sentences and spontaneous talk) and two read speech tasks (reading 55 phonetically and phonologically rich sentences and reading three short passages). KSS DB includes a 16-bit, 44.1kHz speech waveform file and a orthographic file for each speech task.

Keywords: Korean Standard Speech Database, speech corpus, read speech, spontaneous speech, speaker identification

1. 서론

이 연구는 신지영 외(2015)의 후속 연구로, 대검찰청에서 발주한 '용의자 음성식별을 위한 한국인 음성 데이터베이스 수집 및 음성 자동분석 시스템 개발(2014.5.~2014.11.)', '용의자 음성식별을 위한 한국인 표본 음성데이터베이스 구축(2015.4.~2015.11.)', '용의자 음성식별을 위한 한국인 표본데이터베이스 구축2(2016.5~2016.11)' 등 총 3년에 걸쳐 수행한 연구 용역 과제의 결과를 종합 정리하여 학계에 보고하는 것을 목적으로 한다.

신지영 외(2015)에서도 논의하였듯이, 언어 연구를 위해 구축된 음성 코퍼스는 대표성과 균형성을 갖추어야 한다. 한국어 음성 코퍼스에서 대표성이란 모집단인 한국어 사용자들에 대해 일반화가 가능한 정도의 표본 집단의 크기나 자료의 크기를 말한다. 또, 균형성이란 코퍼스를 구성하는 음성 자료의 구성이 한쪽으로 치우치지 않고 고르게 구성되는 것을 의미한다. 그간 여러 연구 기관에서 다양한 목적에 따라 여러 종류의 음성 코퍼스가 구축되었지만, 대표성과 균형성을 모두 갖추었다고 볼 수 있는 코퍼스를 찾기는 어려웠던 것이 사실이다.

* 이 논문은 2016년 대검찰청 연구 용역의 지원으로 수행되었습니다(과제명: 용의자 음성식별을 위한 한국인 표본 데이터베이스 구축 2 , 지원번호: 12168092300).

** 고려대학교, shinjy@korea.ac.kr, 교신저자

Received 8 March 2017; Revised 2 May 2017; Accepted 14 May 2017

본 연구팀이 대검찰청에서 발주를 받아 총 3 년간의 연구를 통해 구축한 '한국인 표준 음성 데이터베이스(이하 KSS DB)'는 한국어의 음성적 특징을 다각적으로 살펴볼 수 있을 만큼의 대표성과 균형성을 설계 시부터 염두에 두었다. 데이터베이스의 구축을 통해 음성학적 연구는 물론, 화자 인식 시스템 개발에 활용될 것을 목표로 설계되었기 때문이다. 이를 위해 전국 단위 총 3,000 명 이상의 발화를 직접 수집 방법과 간접 수집 방법을 통해 데이터베이스화하는 것을 목표로 하였다[1].

이 논문에서 필자는 KSS DB를 구축한 자세한 과정과 그 결과 구축된 DB에 대한 상세한 보고를 함으로써 앞으로 새로운 음성 데이터베이스를 설계하거나 구축하는 데 도움을 줄 수 있을 것으로 기대한다[2].

2. 조사 대상

음성 데이터베이스의 대표성을 확보하기 위해 KSS DB는 조사 대상자(발화자)의 다양한 배경을 고려하였다. 발화의 음성적 특징에 영향을 미칠 수 있는 발화자의 배경으로 가장 크게 고려한 것은 지역, 성별, 연령이었다.

우선 지역은 크게 단일 방언권과 복합 방언권, 그리고 기타로 나누었다. 단일 방언권은 지역 방언적 특성을 고려하여 총 9 개 권역, 즉 수도권, 강원권, 경남권, 경북권, 전남권, 전북권, 제주권, 충남권, 충북권으로 나누었다. **단일 방언권** 화자는 '해당 지역에서 태어나 해당 지역에서 계속 살면서 현재 해당 지역에 거주하고 있는 사람, 혹은 초·중·고 시절 이외의 시기에 타 방언권 거주 경험이 3 년 이내인 사람'으로 정의하고 대상자를 섭외하였다[3].

한편, **복합 방언권** 화자는 단일 방언권 화자와는 달리 경상도와 전라도 지역에서 태어나 자라다가 수도권으로 이주하여 현재 수도권에 거주하고 있는 화자를 의미한다. 복합 방언권은 이주 연수에 따라 두 집단으로 나누었다. 이주한 지 5 년 이상 15 년 미만인 사람을 한 집단으로 하고, 이주한 지 15 년 이상인 사람을 또 한 집단으로 하였다. 대체로 이주한 지 5 년 이상 15 년 미만인 집단은 20 대와 30 대 화자가, 15 년 이상 집단은 40 대 이상 화자가 주조사 대상이 되었다.

그리고 **기타**는 단일 방언권 화자의 조건이나 복합 방언권 화자의 조건에 맞지 않는 화자들을 의미한다. 특별한 경우가 아니라면 단일 방언권과 복합 방언권 화자들을 조사 대상자로 삼았으나 예외적으로 단일 방언권과 복합 방언권 화자에 속하지 않는 화자를 일부 포함하게 되어서 이들을 기타로 분류한 것이다.

기타로 분류된 화자들은 모두 '연속 조사 대상자'들이었다. **연속 조사 대상자**란 여타의 조사 대상자들과는 달리 연구 수행 기간인 3 년 동안 지속적으로 8 회 이상 녹음을 수행한 화자들을 의미한다. 지속적인 녹음을 수행해야 하는 과제의 특성상 지속적인 녹음이 가능한 화자이어야 한다는 조건이 화자 섭외의 최우선 조건이었기 때문에 단일 방언권이나 복합 방언권 외의 화자가 일부 포함된 것이다[4].

성인 화자만을 대상으로 하였기 때문에 20 대, 30 대, 40 대, 50 대, 60 대 이상(60 대+) 등 5 개 연령층을 고려하였다. 그리고 성별은 각 지역과 연령에 대해 남성과 여성의 2 개 집단을 고려하였다.

2.1. 간접 수집 조사 대상자

간접 수집의 경우는 기존의 공개 코퍼스 혹은 연구자의 보유 코퍼스를 연구 목적에 맞게 가공하는 방법으로 이루어졌다.

공개 코퍼스와 연구자 보유 코퍼스 중 KSS DB에 포함될 수 있는 자료를 수집하였다. 발화자의 지역 방언은 자료를 가장 많이 확보할 수 있는 수도권으로 한정하였다. KSS DB에서 활용한 공개 코퍼스로는 국립국어원에서 구축한 '서울말 낭독체 발화'와 SiTEC에서 구축한 '외국인의 한국어 발화 음성 DB' 중 한국어 화자 발화 자료가 포함되었고, 연구자 보유 코퍼스로는 '말글 비교 실험 DB'와 '감정 DB'가 포함되었다. <표 1>은 간접 수집 방법을 활용하여 얻은 자료의 연령별, 성별 구성을 보인 것이다.

표 1. 간접 수집 자료의 연령별 성별 인원(단위: 명)
Table 1. Gender and age group of speakers (indirect methods)

	남	여	합계	%
20대	128	141	269	74.3
30대	30	4	34	9.4
40대	0	20	20	5.5
50대	11	17	28	7.7
60대 이상	9	2	11	3.0
합계	178	184	362	100.0

1 이 논문에서 보고하고 있는 내용은 대검찰청의 연구 용역 수행 내용에 한정되지 않음을 밝혀 둔다. 이 논문에는 대검찰청의 연구 용역 종료 이후 고려대학교 음성언어정보연구실의 자체 재원을 통해 수집되고 가공되고 정련된 내용이 일부 포함되어 있다. 따라서 대검찰청이 납품을 받아 보유하고 있는 DB와 고려대학교 음성언어정보연구실에서 보유하고 있는 DB는 규모와 정련 정도 면에서 일치하지 않는다. 본 논문에서 보고하고 있는 내용은 이 논문의 초고가 작성된 2017년 2월 현재 고려대학교 음성언어정보연구실이 보유하고 있는 DB를 기준으로 하고 있음을 밝힌다.

2 이 논문의 목적은 KSS DB의 특징과 그 구축 과정을 상세히 보고하는 데 있다. 따라서 국내외 음성 코퍼스 현황에 대한 논의는 포함하지 않았다. 국내외 음성 코퍼스의 현황과 선행 연구에 대해서는 신지영 외(2015)를 참고하기 바란다.

3 단일 방언권 조사 대상자를 이와 같이 조작적으로 정의한 배경에 대해서는 신지영 외(2015)에 자세히 논의하였다.

4 뒤에 자세히 논의하겠지만 3년간 총 8회 이상의 녹음을 수행한 '연속 조사 대상자'의 총 수는 81명이었고, 이 가운데 기타로 분류된 화자들은 총 9명이었다.

기존 자료를 활용하였기 때문에 20 대 화자의 자료가 가장 비율이 높았다. 전체 자료의 74.3%가 20 대 화자의 자료였다. 전체적으로 남성과 여성의 비율은 거의 같았지만, 연령에 따라서는 균형이 맞지 않은 경우도 있었다. 30 대 화자의 경우는 남성 화자가, 40 대 화자의 경우는 여성 화자가 압도적인 비율을 보였다[5].

2.2. 직접 수집 조사 대상자

2.2.1. 단일 방언권

직접 수집의 경우는 조사 대상자의 다양한 배경을 충분히 고려하여 설계하였다.

지역 방언 배경과 연령, 성별 등 사회적 변수를 고려하여 표집을 설계할 때는 두 가지 방법이 가능하다. 첫째는 고려하는 모든 변수에 대해 동일한 비율을 배분하는 방법이다. 둘째는 실제 모집단의 비율을 고려하여 표집의 비율을 설계하는 방법이다.

첫 번째 방법은, 전체 목표 인원을, 고려하고자 하는 변수로 나누어 균등 배분하는 방법으로 매우 단수한 방법이다. 하지만 두 번째 방법은 실제 모집단의 비율을 고려하여 표본 집단을 설계하는 방법으로 모집단에 대한 정보를 반영하는 방법이다.

1 차년의 경우는 두 번째 방법, 즉 모집단의 비율을 표집의 비율에 적용하여 설계하되, 최소 인원을 확보하는 방법으로 데이터베이스를 설계하였다(신지영 외, 2015). 하지만 2 차년의 경우는 설계를 달리하여 모든 변수에 대해 가능한 한 동일한 비율을 배분하는 방법을 고려하였다. 그리고 마지막 3 차년에는 다시 인구통계 비례를 고려하여 표집하는 방법으로 조사 대상자의 목표 인원을 설정하였다.

이렇게 연차별로 설계를 달리함으로써, 고려해야 할 모든 변수에 대해 최소 수치를 확보함과 동시에 인구 통계적 비례를 반영할 수 있었다. 이러한 연차별 조정 과정을 통해 설계된 3 년간 목표 인원은 <표 2>에 보인 것과 같다.

표 2. KSS DB 설계 시 지역별, 연령별 직접 수집 조사 대상자 목표 인원(단위: 명. 해당 지역 혹은 연령의 총 인원을 음영으로 표시함)
Table 2. Target number of speakers for each region, gender and age group (direct methods)

	20대	30대	40대	50대	60대	총
수도권	110	125	125	110	80	550
경남권	65	80	80	65	50	340
경북권	65	73	72	58	42	310
전남권	50	58	57	50	35	250
전북권	35	35	35	35	35	175
충남권	50	58	57	50	35	250
충북권	35	35	35	35	35	175
강원권	35	35	35	35	35	175
제주권	35	35	35	35	35	175
총	480	534	531	473	382	2,400

<표 2>에 제시한 목표 인원은 단일 지역 방언 화자를 대상으로 1 년차 600 명, 2 년차 900 명, 3 년차 900 명, 총 2,400 명에 대한 것이다. <표 2>에 제시된 지역과 연령의 목표치는 다음과 같은 방법으로 산출되었다. 우선 1 차년과 3 차년의 목표 인원 총 1500 명에 대해서는 신지영 외(2015)에 제시한 인구 1,000 명당 지역, 연령별 최소 인원 확보를 위한 조정 인구통계를, 1,500 명당으로 환산한 후에 2 차년의 목표 인원 총 900 명을 각 연령대와 지역별로 균등 배분하는 방식으로 더하는 방법이었다. 연구팀이 9 개 지역, 5 개 연령층을 고려하여 설계하였으므로 균등 분배 방식으로 900 명을 나누면 각 지역별로 각 연령층당 20 명씩이 배분된다.

<표 2>와 같은 인원 구성을 목표로 하였으나 직접 조사 대상자의 실제 수집 결과는 <표 3>에 보인 것과 같았다.

<표 3>에 보인 것과 같이 실제 수집 인원은 2,615 명이었다. 이는 목표 인원인 2400 명보다 215 명을 초과하는 수치였다. 지역별 연령별 목표 대비 실제 수집 인원의 차이를 <표 4>에 제시하였다.

표 3. KSS DB의 지역별, 연령별 직접 수집 조사 대상자 실제 수집 인원(단위: 명. 해당 지역 혹은 연령의 총 인원을 음영으로 표시함)
Table 3. Actual number of speakers for each region, gender and age group (direct methods)

	20대	30대	40대	50대	60대+	총
수도권	129	145	145	113	47	579
경남권	131	75	76	80	30	392
경북권	111	56	98	73	29	367
전남권	103	38	69	61	19	290
전북권	110	37	40	59	22	268
충남권	91	39	51	39	29	249
충북권	44	22	37	44	11	158
강원권	76	15	29	22	8	150
제주권	62	40	22	22	16	162
총	857	467	567	513	211	2615

5 간접 조사 대상자에 대한 기술이 신지영 외(2015)의 보고와 약간의 차이가 있는 이유는 논문의 작성 이후에 발화 자료의 재검토 과정에서 일부 오류가 확인되어 수정이 이루어졌기 때문이다.

표 4. KSS DB의 목표 대비 수집 결과의 지역별, 연령별 비교(단위: 명. 양수는 목표보다 실제 수집이 많은 것을, 음수는 목표보다 실제 수집이 적은 것을 의미하며 음영으로 표시함)

Table 4. The difference between target and actual number of speakers for each region, gender and age group (direct methods)

	20대	30대	40대	50대	60대	**총**
수도권	19	20	20	3	-33	**29**
경남권	66	-5	-4	15	-20	**52**
경북권	46	-17	26	15	-13	**57**
전남권	53	-20	12	11	-16	**40**
전북권	75	2	5	24	-13	**93**
충남권	41	-19	-6	-11	-6	**-1**
충북권	9	-13	2	9	-24	**-17**
강원권	41	-20	-6	-13	-27	**-25**
제주권	27	5	-13	-13	-19	**-13**
총	**377**	**-67**	**36**	**40**	**-171**	**215**

<표 4>에 제시한 것과 같이 전체 9 개 권역 중 5 개 권역은 목표를 초과하였고, 4 개 권역은 목표에 미달하였다. 하지만 목표에 미치지 못한 수가 큰 규모는 아니었다. 목표에 비해 조사 인원이 가장 적었던 지역은 강원권으로 총 25 명이 목표에 미달하였다. 다음은 충북권과 제주권으로 각각 17 명과 13 명의 인원이 목표에 미달하였다. 그리고 충남권의 경우는 1 명이 목표에 미달하였을 뿐이었다. 한편, 5 개 지역은 목표 인원을 초과하여 수집되었는데, 초과 인원이 가장 많은 지역은 전북권으로 93 명이 목표를 초과하였다. 다음은 경북과 경남권에서 각각 57 명과 52 명의 화자가 목표를 초과하여 수집되었다. 그리고 전남권 40 명, 수도권 29 명의 화자도 초과하여 수집되었다.

연령층별로 목표 인원 대비 수집 인원을 비교해 보았을 때 가장 목표에 미치지 못했던 연령층은 60 대 이상이었다. 60 대 이상 조사 대상자는 목표 대비 전 지역에서 수집 인원이 부족하였다. 총 171 명이 목표에 미치지 못하였다. 다음으로 목표 대비 인원이 부족한 연령층은 30 대로 수도권과 전북권, 제주권 등 3 개 지역을 제외한 6 개 지역에서 목표에 미치지 못하는 조사가 이루어졌다. 반면에 나머지 20 대, 40 대, 50 대의 경우는 전체적으로 목표를 초과하는 조사 대상에 대한 녹음이 이루어졌다. 특히 20 대는 전 지역에서 목표 대비 초과 실적이 두드러지게 높은 연령층이었다.

다음은 조사 대상자의 성별 비율을 정리하였다. <표 5>는 KSS DB의 남녀 비율을 보인 것이다. 표에 보인 바와 같이 전 연령층에서 여성 조사 대상자가 남성 조사 대상자보다 높은 비율을 보였다. 전체적으로는 남녀의 비율이 4:6 정도를 보이는 것으로 나타났다. 특히, 40 대와 50 대에서 여성 조사 대상자의 비율이 높았다.

표 5. KSS DB 성별 비율(단위: %. 해당 지역 혹은 연령의 총 성별 비율은 음영으로 표시함.)

Table 5. The proportion of male and female speakers in KSS DB

	20대		30대		40대		50대		60대+		지역	
	남	여	남	여	남	여	남	여	남	여	남	여
수도권	50	50	43	57	28	72	33	67	32	68	38	62
경남권	46	54	51	49	32	68	38	63	50	50	43	57
경북권	37	63	36	64	41	59	34	66	38	62	37	63
전남권	49	51	37	63	22	78	41	59	37	63	38	62
전북권	39	61	38	62	28	73	56	44	59	41	43	57
충남권	48	52	38	62	53	47	44	56	59	41	48	52
충북권	45	55	36	64	27	73	39	61	36	64	37	63
강원권	36	64	33	67	28	72	50	50	50	50	37	63
제주권	21	79	73	28	41	59	41	59	50	50	42	58
연령	42	58	44	56	32	68	40	60	45	55	40	60
	남	여	남	여	남	여	남	여	남	여	남	여
	20대		30대		40대		50대		60대		지역	

2.2.2. 복합 방언권 및 기타

복합 방언권이란 앞서도 언급하였듯이 경상도와 전라도 지역에서 태어나 자라다가 수도권으로 이주하여 현재 수도권에 거주하고 있는 화자 집단을 말한다. 이주 연수에 따라 5년 미만과 15년 이상의 화자 집단으로 세분하였다. 복합 방언권 화자 집단은 1차년에는 고려하지 않았고, 2차년과 3차년에 각 100명씩 총 200명을 수집 목표로 설정하였다. 지역은 화자의 원 방언권에 따라 경상권과 전라권의 두 지역으로 한정하였다. 복합 방언권이 경우에는 연령과 성별에 대한 구체적인 목표를 설정하지는 않았다.

2 차년과 3 차년 연구를 통해 수집한 복합 방언권 화자는 총 205 명이었다. 지역적으로는 경상권 화자가 122 명, 전라권 화자가 83 명이었다. 연령은 20 대 화자가 가장 많아서 총 95 명이었고, 30 대 40 명, 40 대 28 명, 50 대 35 명, 60 대 이상 7 명이었다. 조사 대상자의 성별을 살펴보면 남성 화자가 97 명, 여성 화자는 108 명으로 여성 화자의 비율이 조금 높았다.

한편, 기타로 분류된 화자는 화자의 지역 방언 배경이 단일 방언권의 기준에도, 복합 방언권의 기준에도 들지 못하는 경우를 말한다. 방언 배경보다는 3 년간 8 회 이상의 지속 녹음 가능 여부가 우선으로 고려된 결과였다. 기타에 속하는 화자는 총 2,829 명의 직접 조사 대상자 중 9 명에 불과했다.

<표 6>은 복합 방언권과 기타로 분류된 조사 대상자들의 지역, 연령과 성별 구성을 정리한 것이다.

표 6. KSS DB 복합 방언권과 기타의 연령별 성별 수집 인원(단위: 명. 해당 지역 혹은 연령의 성별 합계와 총 합계는 음영으로 표시함)
Table 6. The number of speakers of complex dialectal background and other dialectal background for gender and age group

		20대		30대		40대		50대		60대+		지역	
		남	여	남	여	남	여	남	여	남	여	남	여
복합	경상	43	27	12	14	7	4	4	7	0	4	66	56
	전라	13	12	6	8	6	11	4	20	2	1	31	52
연령		56	39	18	22	13	15	8	27	2	5	97	108
		95		40		28		35		7		205	
기타		1	1	2	0	1	1	1	2	0	0	5	4
연령		2		2		2		3		0		9	
		남	여	남	여	남	여	남	여	남	여	남	여
		20대		30대		40대		50대		60대		지역	

2.2.3. 연속 조사 대상자

연속 조사 대상자란 연구 과제가 진행된 3년 동안 8회 이상의 녹음을 주기적으로 수행한 조사 대상자들을 말한다. 동일 화자의 발화를 일정한 간격으로 주기적으로 녹음함으로써 음성적 특징 중 변화하는 요소와 불변화하는 요소를 알아보기 위해 설계된 화자 집단이었다. 100명이 대상자로 섭외되어 녹음을 시작하였으나, 8회 이상의 발화 녹음을 유지한 조사 대상자는 총 81명이었다. 따라서 녹음 유지율은 81%였다.

연속 조사 대상자의 지역별 분포는 다음과 같았다. 수도권 45 명, 경북권 14 명, 복합 방언권 13 명, 기타 방언권 9 명이었다. 연령별로는 20 대가 30 명, 30 대가 30 명, 40 대가 12 명, 50 대가 8 명, 60 대가 1 명이었다. 한편, 성별로는 남성이 47 명, 여성이 34 명이었다. 조사 대상자의 지역, 연령, 성별 상세 정보는 <표 7>에 정리하였다.

표 7. KSS DB 연속 조사 대상자 성별 수집 인원(단위: 명. 연속 조사 대상자란 연구 과제 수행 기간 중 8회 이상 주기적으로 수행한 자를 의미한다. 해당 지역 혹은 연령의 성별 합계와 총 합계는 음영으로 표시함)
Table 7. The number of speakers of region, gender, and age group recording more than eight times for project period

	20대		30대		40대		50대		60대+		지역	
	남	여	남	여	남	여	남	여	남	여	남	여
수도권	8	12	7	13	1	1	0	2	1	0	17	28
경북권	1	1	0	1	2	6	1	2	0	0	4	10
복합	5	1	3	4	0	0	0	0	0	0	8	5
기타	1	1	2	0	1	1	1	2	0	0	5	4
연령	15	15	12	18	4	8	2	6	1	0	34	47
	30		30		12		8		1		81	

지금까지 논의한 바와 같이 KSS DB의 총 조사 대상은 간접 수집 363 명, 직접 수집 2,829 명으로 총 3,192 명이었다. 3,192 명의 조사 방법별, 지역별, 연령별, 성별 정보 등 자세한 내용은 <부록 1>에 실었다[6].

3. 자료 수집 방법 I: 간접 수집

간접 수집이란 공개된 코퍼스나 연구자가 보유하고 있는 코퍼스 중에서 KSS DB의 구축 방향과 목적에 맞는 음원을 수집하여 연구에 활용할 수 있도록 데이터베이스화한 것을 의미한다. 간접 조사를 통해 KSS DB에 포함된 자료는 다음과 같은 기준으로 선정하였다. 첫째, 발화자의 지역, 연령, 성별 정보가 정확히 확인될 수 있는 자료일 것, 둘째 직접 수집과 같은 방식으로 발화한 자료, 즉 발화 겹침 등이 없고 단독으로 발화된 자료일 것, 셋째 한 화자의 발화의 양이 일정 규모 이상[7]이어야 할 것 등이었다.

이러한 기준을 충족시키는 공개 자료 중에는 국립국어원에서 구축한 서울말 낭독체 발화 DB와 SiTEC에서 구축한 외국인 한국어 발화 음성 DB 중 한국인 발화 자료가 포함되었다.

서울말 낭독체 발화는 2 대 이상 서울 경기 지역에 거주해 온 다양한 연령대의 서울말 화자 120 명이 소설, 수필, 논설 등에서 뽑은 다양한 형태의 글을 낭독한 음성 자료를 DB화한 것이다. 연령에 따라 낭독 분량의 차이가 있었는데, 20 대 ~ 40 대 화자는 930 문장을 낭독하였고, 50 대 이상의 화자는 404 문장을 낭독하였다. 발화 스타일은 낭독체였고, 낭독 방식은 화자 스스로가 정하되, 스스로에게 가장 편안한 속도와 크기로 수행는 방식이었다. 총 120 명이 녹음에 참여하였으나, 실제 배포된 데이터베이스를 통해서는 118 명의 녹음 파일만 활용할 수 있었다.

외국인의 한국어 발화 음성 DB는 한국어 학습자들의 발화를 수집하는 것을 목적으로 구축된 것이지만, 그 가운데 외국인 화자와의 대조를 위해 한국이 남녀 10 명의 발화 자료가 포함되어 있다. 화자는 20 대 ~ 30 대 화자였고 학습자들의 발화 특성을 알아보기 위해 설계된 단문 10 개, 대화문 20 개, 단어 88 개를 낭독한 것이다. 대화문이기는 하지만 두 사람의 대화가 아니라 대화의 지문을 한 사람이 낭독한 것이라 KSS DB의 일부로 포함하였다.

연구자가 보유한 코퍼스 중에는 말글 비교 실험 DB와 감정 DB를 포함하였다. 우선 말글 비교 실험 DB는 20~30 대 화자 204 명이 세 가지 말하기 과제 중에서 두 가지 과제를 수행하며 각 과제당 3~5 분 정도의 발화를 수행한 것이다. 말하기 과제는 면접 상황의 일방적 말하기, 방송 보도 상황, 그리고 수업 발표 상황이었다.

또한, 감정 DB는 4 가지 감정(평상, 기쁨, 화남, 슬픔)을 표현할 수 있는 대화 형식의 문장을 발화한 자료다. 대화 형식의 대본이었으나 발화자의 단독 발화만 DB화되어 있었기 때문에 수집 대상으로 삼았다. 총 42 명의 화자의 발화가 수집되어 있었지만, 전문가 집단의 12 명 발화분은 KSS DB에 포

6 논문에 보고된 수집 인원은 과제가 종료된 이후, 고려대학교 음성언어정보 연구실의 재원으로 수집된 인원을 포함한다는 점을 밝힌다.
7 한 사람의 발화 분량이 5분 이상인 것을 우선 고려 대상으로 삼았다.

함하지 않았다. 따라서 KSS DB에 포함된 자료는 일반인 30 명의 발화였다. 각 화자들은 다양한 종결법을 가진 65 개 문장 혹은 40 개 문장을 4 가지 감정으로 2 회 발화한 자료였다.

대상이 되는 4 종의 코퍼스 내 음성을 일일이 확인하는 방법으로 KSS DB에 포함할 자료를 수집하였다. KSS DB는 직접 수집 방법을 주된 수집 방법으로 설계되었기 때문에 간접 수집 방법은 1 차년에만 수행되었다.

(1)은 간접 수집 방법으로 KSS DB에 포함된 자료를 개관한 것이다.

(1) a. 공개 코퍼스
가. 서울말 낭독체 발화
- 기관: 국립국어원
- 내용: 소설, 수필, 논설 등에서 발췌한 문장
- 발화 분량
20 대 ~ 40 대 화자: 930 문장
50 대 이상 화자: 404 문장
- 발화 스타일
낭독체 발화
- 인원: 20 대 ~ 60 대 남녀 화자 총 118 명

나. 외국인의 한국어 발화 음성 DB
- 기관: SiTEC
- 내용: 단문 10 개, 대화문 20 개, 단어 88 개 낭독
- 인원: 한국인 남녀 화자 10 명

b. 연구자 보유 코퍼스
가. 말글 비교 실험 DB
- 내용: 2 가지 종류의 말하기 과업 수행
- 인원: 20·30 대 남녀 화자 총 130 명(2008 년)
20·30 대 남녀 화자 총 74 명(2010 년)

나. 감정 DB
- 내용: 3 가지 종결법의 총 65 개 문장을 4 가지 감정으로 2 회씩 발화(2005 년)
3 가지 종결법의 총 40 개 문장을 4 가지 감정으로 2 회씩 발화(2006 년)
- 인원: 20 대 남녀 화자 각 5 명, 총 10 명(2005 년)
20 대 남녀 화자 각 10 명, 총 20 명(2006 년)

간접 수집 방법으로는 400 명의 발화 수집을 목표로 하였으나, 실제 수집하여 구축된 DB에는 362 명의 발화가 포함되었다.

4. 자료 수집 방법 II: 직접 수집

KSS DB의 주력 자료 수집 방법은 직접 수집 방법이었다. 직접 수집 방법으로 대규모 자료를 수집하기 위해서는 구체적인 수집 프로토콜을 개발해야 한다. 이를 위해 직접 수집의 절차를 크게 녹음 전 단계, 녹음 단계, 녹음 후 단계의 세 단계로 나누어 각 단계에서 고려할 사항들을 정리하였다.

녹음 전 단계에는 조건에 맞는 조사 대상자 섭외, 녹음 장소 확보, 녹음 장비를 점검, 섭외된 조사 대상자에 대한 필요 서류 작성, 녹음 과정 설명 등이 포함된다. 우선 녹음 장소는 녹음실을 확보하는 것이 좋지만 조사 상황을 고려할 때 녹음실을 확보하지 못하는 경우가 대부분일 것으로 예상되었기 때문에 최대한 조용하고 울림이 없는(혹은 적은) 공간을 확보할 수 있도록 유의하였다.

조사 대상자로서의 조건에 맞는 조사 대상자의 섭외를 위해서는 조사 대상자에 대한 사전 정보가 필요하다. 하지만 지역, 연령대, 성별에 대한 사전 정보 외에도 더 구체적이고 자세한 조사 대상자에 대한 정보가 필요하기 때문에 녹음 전 단계에서 조사 대상자가 직접 피험자 정보표(<부록 2> 참조)를 작성하는 것이 필요하다. 이와 더불어 법적인 문제를 해결하기 위해 조사 대상자에게 '피험자 동의서' 작성을 요구하는 절차도 밟아야 한다.

피험자 정보표와 피험자 동의서 작성을 마치면, 녹음 단계에 들어가게 된다. 녹음은 본 연구를 위해 설계된 5 가지 발화 과제를 차례로 수행하는 방법으로 이루어진다. 조사원은 조사 대상자가 녹음하는 동안 녹음 내용이 잘 수행되고 있는지 모니터링을 해야 한다.

녹음을 마치면 녹음 후 단계에 들어가게 된다. 녹음 후 단계에서 조사원은 조사 대상자들에게 소정의 사례비를 지급하고 '녹음비 수령 영수증'을 발급하며, '녹음비 수령자 목록'을 작성한다. 직접 수집을 위한 전체 과정을 거치는 데 소요되는 시간은 피험자 1 인당 약 30 분 내외였다. <그림 1>은 직접 수집의 세 단계를 정리한 것이다.

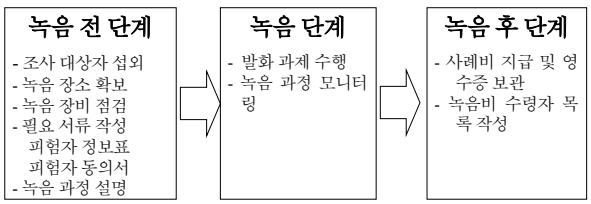


그림 1. 직접 수집의 세 단계
Figure 1. Three steps for collecting speech samples

4.1. 조사원 교육 및 조사원 파견

효율적인 자료 조사를 위해 연구팀의 연구원 외에도 전국 단위로 조사원을 모집하고 이들을 교육시킨 후에 조사에 투입하였다. 조사원 모집은, 1) 지역 대학 관련 전공 교수들의 추천, 2) 해당 지역 지인들의 추천, 3) 인터넷 공지 등등의 방법으로 이루어졌다.

처음에는 각 지역의 관련 전공 학부생 혹은 대학원생을 조사원으로 모집하고 조사 활동을 수행하도록 하였으나, 대학생이나 대학원생들의 경우는 다양한 연령층의 조사 대상자, 특히 40 대 이상의 조사 대상자를 섭외하는 데 어려움이 있었다. 이를 보완하기 위하여 중장년층 조사원을 확보하기 위해 노력했다. 실제로 중장년층의 조사 대상자 섭외를 위해서는 중장년층 조사원을 활용한 것이 효과적이었다.

신지영 외(2015)에도 논의하였듯이 대학생이나 대학원생 등 학생 조사원과 중장년층 조사원들은 서로 다른 특성을 갖는다. 학생 조사원의 경우는 음성 파일의 처리나 전사 등 후처리 작업을 다양하게 수행할 수 있는 장점이 있지만 자신의 또래 집단 이외의 다양한 연령층의 조사 대상자 섭외를 어려워했다. 반면에 중장년층 조사원의 경우는 다양한 연령층의 조사 대상자 섭외에 뛰어났지만, 연구팀이 원하는 수준의 후처리를 수행하는 데에는 어려움이 있었다. 이에 연구팀은 학생 조사원의 섭외 한계는 중장년층 조사원으로 보완하고, 중장년층 조사원의 후처리 수행의 한계는 연구팀이 보완하는 전략을 사용하였다.

모집된 조사원들은 모두 조사에 필요한 교육을 받았다. 조사의 일관성을 위해 본 연구팀은 3 시간 정도의 조사원 교육 프로그램을 개발하였다[8]. 연구팀이 개발한 오프라인 교육 프로그램을 수강한 사람들만 조사원으로 활용하였다.

조사원 교육 프로그램은 모두 네 가지 내용으로 구성하였다. 첫 번째 내용은 조사 목적과 조사 내용에 대한 것으로, 조사원들에게 본 연구의 조사 목적과 내용을 상세히 설명하는 내용이다. 이를 통해 본 연구를 위한 조사 대상자의 섭외 조건을 숙지하는 것을 목표로 하였다. 두 번째 내용은 조사 방법에 대한 것으로, 구체적인 조사 방법을 상세히 설명하는 내용이다. 이 과정을 통해 조사원들이 자료 수집의 세 단계에 대해 숙지할 수 있도록 하였다. 세 번째 내용은 녹음기 사용법을 교육시키는 것이었다. 이를 통해 조사원들이 자신들이 조사에서 사용하게 될 디지털 녹음기의 작동 방법에 익숙해질 수 있도록 하였다. 마지막 네 번째 내용은 조사 및 전사 실습이었다. 이를 위해 조사원들끼리 조사원과 조사 대상자로 역할을 번갈아 맡아 조사를 직접 수행하는 실습을 하였다. 연구원들은 조사원 교육을 받고 있는 교육생들이 수행하는 실습의 전 과정을 면밀히 관찰한 후에 개별 교육생들의 수행 내용에 대한 피드백을 주었다.

이러한 과정을 거쳐 조사원 교육을 받고 조사에 투입된 조사원은 1 차년 22 명, 2 차년 53 명, 3 차년 55 명으로 총 124 명이었다[9]. 이 가운데는 3 년을 모두 조사원으로 활동한 사람이 2 명, 2 차년과 3 차년 총 2 년 동안 조사자로 활동한 사람이 2 명 포함되어 있었다. 나머지 조사원들은 조사 대상자 섭외의 한계 등을 이유로 1 년 동안만 조사원으로 활동하였다.

4.2. 수집 자료 설계

직접 수집의 과정을 통해 수집하고자 하는 자료의 설계는 다음의 세 가지 조건을 고려하여 수행하였다. 첫째, 조사 대상자 1인당 최소 10분 이상의 발화를 수집한다. 둘째, 대본을 활용한 낭독 발화와 대본 없이 발화자가 발화하는 자유 발화의 두 가지 유형의 발화 과제가 모두 포함되도록 설계한다. 셋째, 20세 이상의 전 연령층이 조사 대상자이므로 모든 연령층의 조사 대상자들이 큰 어려움 없이 수행할 수 있는 과제가 될 수 있도록 과제를 설계한다.

이러한 세 가지 기준을 가지고 크게 낭독 발화 과제와 자유 발화 과제에 속하는 발화 과제를 설계하였다. 낭독 발화는 연구팀의 연구 목적에 맞게 설계된 대본을 대본에 적힌 대로 조사 대상자가 읽도록 하는 것을 의미한다. 따라서 모든 조사 대상자자로부터 동일한 내용의 음성을 얻기에 유용한 방법이다. 한편 자유 발화는 발화의 내용이 조사 대상자마다 달리 실현되는 것으로, 발화 수행이 자연스러운 장점을 가지며 조사 대상자의 음성적 특징을 최대한 반영하는 데 유용한 방법이다.

본 연구에서는 이와 같은 각 발화 유형의 장점을 모두 포괄하기 위하여 <그림 2>에 보인 것과 같이 모두 5 가지 발화 과제를 설계하였다.

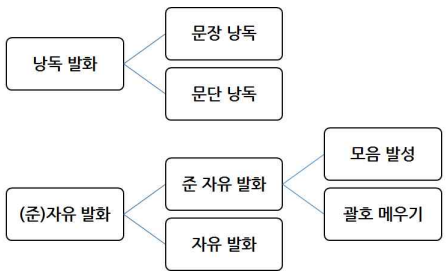


그림 2. 직접 수집 발화 과제 설계
Figure 2. Design of speech tasks for collecting speech samples

그림에 보인 바와 같이 우선 발화 과제는 낭독 발화 과제와 자유 발화 과제로 나누어 구성하였다. 낭독 과제는 다시 두 가지 유형의 과제로 나누어 구성하였는데, 문장 낭독 과제와 문단 낭독 과제가 그것이었다.

문장 낭독 과제는 한국어의 모든 음운이 고루 실현되도록 개발한 55 개 문장을 낭독하는 과제로, 55 개 문장은 한국어 분절음의 특성을 수집하기 위해 설계된 문장 39 개와, 특징적인 음운 현상을 관찰하기 위해 설계된 문장 16 개로 구성하였다.

한국어 분절음을 조사하기 위해 개발된 39 개 문장은 목표 분절음(자음 18 개, 단모음과 이중모음 21 개)이 문장의 시작에 나타나고, 목표 분절음이 해당 문장에 최소한 3 회 이상 실

8 실제 조사원 교육의 소요 시간은 1시간 30분 정도였다. 조사원 교육은 본 연구팀의 소속 대학에서 진행하는 것을 기본으로 하였다. 하지만 상황에 따라서 일부 조사원에 대한 교육은 본 연구팀의 소속 연구원이 소재지에 가서 진행하기도 하였다. 조사원 교육을 위해 본 연구팀 소속 대학으로 출장을 오게 된 조사원들에게는 출장비를 제공하였다.
9 조사원 교육 이수자 수는 총 130명이었지만, 3차년 조사원 교육 이수자 중에서 6명이 조사를 포기하여 실제로 조사에 투입된 조사원의 수는 124명이었다.

현되도록 설계하였다. 또, 운율과 음운 현상을 관찰하기 위해 개발된 16 개 문장은 지역, 연령, 성별 등에 따라 차이를 보인다고 선행 연구에 보고된 운율 및 음운 현상(예를 들어 성조, 장단, 자음군 단순화, /ㄴ/-/ㄹ/ 연쇄의 발음, 경음화, /ㄴ/ 첨가 등)을 관찰할 수 있는 단어가 다수 포함될 수 있도록 설계하였다. 낭독 문장의 설계 기본 원칙은 (2)에 정리한 바와 같다.

(2) 문장 설계 기본 원칙
a) 모음 문장
- 목표 모음(단모음과 이중 모음)이 어두 위치를 포함하여 3 회 이상 실현
- 어두 위치에 실현된 목표 모음은 단독 음절로 실현
- 비어두 위치에 실현된 목표 모음은 각각 초성에 /ㄱ/와 /ㅎ/[10]를 동반하고 개음절에서 실현
- 어두 위치의 목표 모음에 후행하는 음절의 초성은 양순음, 경음, 격음이 아닌 자음으로 실현

b) 자음 문장
- 목표 자음이 어두 위치를 포함하여 3 회 이상 실현
- 어두 위치에 실현된 목표 자음은 모음 /ㅏ/와 함께 개음절에서 실현
- 비어두 위치에 실현된 목표 자음은 각각 모음 /ㅜ/와 /ㅣ/와 함께 개음절에서 실현

c) 운율 및 음운 현상
- 성조와 장단의 실현, 자음군 단순화, /ㄴ/-/ㄹ/ 연쇄의 발음, 경음화, /ㄴ/ 첨가, 격음화 등 선행 연구를 통해 지역, 연령, 성별에 따라 차이를 보이는 운율 및 음운 현상이 관찰될 수 있는 단어들이 최대한 많이 포함될 수 있도록 문장 구성

문장 낭독 대본은 1 차년 첫 개발 후 총 6 차에 걸쳐 수정이 이루어졌다. 첫 번째 수정은 1 차년 후반기에 이루어졌다. 1 차 수정 시 낭독 문장 중 3 개 문장이 수정되었고, 문장의 제시 순서 일부가 변경되었다. 2 차년에는 1 차년 수집 과정을 통해 발견된 문제점들을 반영하여 두 번째 수정이 이루어졌다[11]. 두 번째 수정 시에는 낭독 문장의 대폭적인 수정이 이루어졌다. 그리고 이후에도 자료를 수집하는 과정에서 발견된 문제점들을 반영하여 3 차 수정(/ㅅ/, /ㅆ/ 문장의 수정)과 4 차 수정(/의/ 문장 추가) 및 5 차 수정(모음 문장 일부 수정), 6 차 수정(일부 문장 수정)이 이루어졌다. 문장 낭독 대본의 최종본은 7 차 버전(6 차 수정본)이었고, 그 완료 시기는 2015 년 4 월 30 일이었다. 따라서 2 차년 초반기 일부 자료를 제외하고는 동일한 문장 낭독 대본을 바탕으로 수집하였다.

한편, 문단 낭독 과제는 문장보다 큰 단위에서 관찰되는 운율적 특징을 관찰하기에 적합한 자료를 수집하기 위해 설계된 것이다. 운율적 특징과 함께, 음성학적, 음운론적 특징을 관찰할 수 있는 3 개의 문단을 개발하였다. 이 가운데 2 개의 문단은 운율적 특징을 잘 관찰할 수 있도록 공명음의 비율을 높였고, 하나의 음운구 단위가 3-4 음절로 자연스럽게 실현될 수 있도록 문장을 구성하였다. 그리고 나머지 한 문단은 일상적이고 친숙한 내용을 쉽게 낭독할 수 있도록 설계하였다. 문단 낭독 대본은 개발된 이후 수정되지 않았다.

한편, 자유 발화 과제는 크게 두 가지 유형의 과제로 구성하였다. 첫 번째 유형은 준 자유 발화 유형이었다. 준 자유 발화 유형이란 대본을 낭독하는 것이 아니므로 낭독 발화 유형에는 속하지 않지만, 주어진 조건이나 자료를 바탕으로 하여 발화를 한다는 점에서 완전 자유 발화라고 하기 어려운 유형의 발화 과제를 말한다. 이러한 과제의 특성상 준 자유 발화 과제라는 이름으로 분류하였다. 두 번째 유형은 자유 발화 유형으로 대본이나 틀이 전혀 주어지지 않은 상태에서 조사 대상자의 자발적인 발화 내용으로 전체 발화를 수행하게 하는 과제였다.

준 자유 발화 과제는 다시 모음 발성 과제와 괄호 메우기 과제의 두 가지 유형으로 설계하였다. 첫 번째 유형인 모음 발성 과제는, 조사 대상자들에게 자신에게 가장 편안한 음높이와 크기로 모음을 3 초 정도씩 연장 발성하여 3 회 반복하도록 하는 과제를 말한다. 모음 연장 발성은 지역, 연령, 성별의 음성적 특징을 살펴보는 데 유용한 자료로 활용될 수 있도록 설계된 것이다. 1 차년과 2 차년의 초반부에는 /아, 이, 우/ 3 개 모음만 연장 발성하게 하였으나, 그 이후에는 모음 글자 'ㅏ, ㅐ, ㅓ, ㅔ, ㅗ, ㅜ, ㅡ, ㅣ' 8 개에 대해 연장 발성을 하도록 수정하였다.

두 번째 유형의 준 자유 발화 과제는 괄호 메우기 과제였다. 괄호 메우기 과제는 "제 이름은 (　)입니다."와 같이 괄호가 포함된 문장을 대본으로 구성하여 피험자들이 자신의 상황에 맞게 세부사항을 직접 채워 넣으며 말하도록 설계한 과제를 말한다. 괄호 메우기 과제는 괄호 안의 내용을 조사 대상자와 관련된 내용으로 구성함으로써 전형적인 낭독과는 달리 훨씬 자연스러운 발화를 유도할 수 있다. 혼자 아무런 대본 없이 일정 시간 이상 발화하도록 요청하는 경우 대부분의 조사 대상자들은 무슨 말을 해야 할지 부담을 느낀다. 따라서 충분한 발화를 유도해 내기 매우 어렵다. 이러한 점에서 괄호 메우기 방식의 과제는 비록 완전한 자유 발화는 아니지

10 초성을 /ㄱ/와 /ㅎ/로 한정한 이유는 혓몸이 조음에 관여하는 만큼, 모음의 조음에 가장 큰 영향을 미칠 수 있는 연구개음의 조건과 가장 적은 영향을 미치는 후두음의 조건을 고려하기 위해서였다.

11 문장 낭독을 위해 설계된 문장은 (2)에 보인 다양한 원칙을 지켜야 하는 만큼, 해당 단어들을 조합하여 자연스러운 문장을 만드는 일이 쉽지 않다. 연구팀은 최대한 자연스러운 문장을 만들기 위해 노력하였으나, 1차년 발화 수집 과정에서 발화자들이 일부 문장의 낭독을 어려워하거나 자주 틀리는 경우를 발견하게 되었다. 또, 문장의 제시 순서에 있어서도 길이가 짧거나 더 부드럽게 읽히는 문장을 앞에 제시하는 것이 낭독에 부담을 줄일 수 있다고 판단되었다. 이러한 문제점을 보완하기 위해 2차년이 수집이 시작되기 전에 낭독 문장에 대한 대대적인 수정이 이루어지게 된 것이다.

만 자발적인 발화에 가까운 발화를 짧은 시간 안에 원하는 분량만큼 얻을 수 있다는 장점을 지닌다. 뿐만 아니라 모든 조사 대상자들에게 유사한 주제와 내용의 발화를 자연스러운 운율로 수집할 수 있다.

본 연구팀에서 개발한 괄호 메우기 과제는 그 주제를 '신상, 가족/친구, 지역/교통, 여가/문화, 상식'의 5 가지로 나눈 후, 각 주제당 4~6 개 문장의 발화가 수행될 수 있도록 구성하였다. 괄호 메우기 과제의 개발 문장은 개발 이후 변경하지 않았다.

한편, 자유 발화 과제는 일방적 말하기 상황에서 조사 대상자가 3~5 분 정도의 이야기를 자유롭게 수행하도록 하는 과제였다. 기본적으로는 피험자가 주제를 자유롭게 선택하도록 요청하였다. 하지만 조사 대상자가 이야기의 소재를 찾는 데 어려움을 느끼는 경우를 대비하여 가족 소개, 아끼는 물건, 지난 주말에 한 일, 좋아하는 영화의 줄거리 소개 등 쉽게 이야기를 이어갈 수 있는 소재를 준비해 두었다. 준비한 이야기 소재를 제공하였음에도 불구하고 이야기를 이어가기 어려워하는 경우를 대비하여 그림 보고 설명하기 과제를 준비해 두었다. 그림 자료(전래 동화의 줄거리를 표현하는 4 단 그림, 가족사진 등)를 준비하여 그림을 보고 설명하며 자유 발화를 수행할 수 있도록 하였다.

수집 자료 전체에 대한 구체적인 내용은 <부록 3>에 실었다.

4.3. 자료 수집 절차

조사 대상자로부터 발화 자료를 수집하는 절차는 다음과 같았다. 우선 조사 대상자가 녹음을 시작하면 총 5가지의 발화 과제를 다음 <그림 3>과 같은 순서로 제시하여 수행하게 하였다.

그림 3. 발화 과제 수행 순서
Figure 3. The order of performing speech tasks

조사 대상자들은 녹음 단계에서 모음 발성을 가장 먼저 수행하도록 하였다. 다음은 문단 낭독, 문장 낭독, 괄호 메우기, 자유 발화의 순서였다. 과제의 제시 순서는 조사 대상자들의 심리적인 부담이 가장 낮은 것으로 시작하여 점차 높아지도록 설계하였다.

조사원은 자료 수집 과정에서 발화자의 발화 수행을 모니터링하도록 교육하였다. 이를 통해 조사원이 조사대상자가 과제를 잘 수행하는지를 녹음 과정 내내 모니터링하는 역할을 수행하도록 하였다. 단, 조사 대상자의 발화 오류에 대한 대처는 과제마다 조금씩 달리 대응하도록 교육하였다. 각 과제별로 오류에 대처 방법을 정리하면 다음과 같다.

우선 모음 발성에서 오류가 생기는 경우는 전체를 다시 수행하도록 하였다. 하지만 모음 발성의 경우는 오류가 거의 나지 않았다. 문단 낭독의 경우는 기본적으로 대본과 조금 다르게 읽는 것을 허용하였다. 또, 문단을 낭독하다가 스스로 수정하여 읽은 경우에도 해당 문단 전체를 다시 읽게 하지 않았다. 하지만 낭독이 부자연스럽거나 낭독이 대본과 너무 많이 달라지거나 낭독 과정에서 너무 자주 틀린 경우에는 해당 문단을 표시해 두었다가 문단 낭독 과제를 모두 마친 후에 해당 문단을 다시 낭독하게 하였다.

한편, 문장 읽기의 경우는 문장에 포함된 분절음들이 관심의 대상이므로 대본과 다르게 읽지 않도록 유의해야 한다. 따라서 조사 대상자가 대본과 달리 읽은 경우는 해당 문장을 표시해 두었다가 대본대로 발화하도록 다시 낭독을 요청하였다. 단, 조사 대상자가 전체 문장을 모두 낭독한 후에, 잘못 발화한 문장을 추려서 다시 낭독하도록 요청하였다. 문단 발화의 경우와는 달리 문장 낭독 중 문장의 일부를 스스로 수정하여 읽은 문장 또한 재낭독의 대상에 포함하였다.

괄호 메우기 과제의 경우는 내용에 대한 발화자의 가감을 어느 정도 허용하는 방향으로 자료를 수집하였다. 마지막으로 수행된 자유 발화의 경우는 3 분 이상의 연속 발화를 수집하는 것을 목표로 하였다. 만약 조사 대상자의 발화 분량이 3 분에 미치지 못하는 경우에는 부족한 발화의 양을 보충하기 위해 연구팀이 준비한 그림 보고 말하기 등을 추가 과제로 하여 발화 분량을 확보하였다.

앞서 언급한 바와 같이 발화 과제의 설계 시 전 연령층의 조사 대상이 수행에 큰 어려움을 느끼지 않을 수 있도록 유의하였다. 하지만 1 차년 자료를 수집하는 과정에서 50 대 이상의 경우 특히 낭독 과제에서 어려움을 느끼고 소요 시간 또한 긴 것을 확인할 수 있었다. 이를 고려하여 1 차년 중간부터는 조사 대상자를 두 그룹(즉 40 대 이하와 50 대 이상)으로 나누고 부여하는 과제의 양에 차등을 두었다.

5 개의 과제 중에서 모음 발성 과제와 문장 낭독 과제, 자유 발화 과제의 경우는 두 연령 집단 사이에 분량의 차이를 두지 않았다. 모음 발성 과제와 자유 발화 과제의 경우는 연령이 높아도 수행에 큰 어려움을 보이지 않았기 때문에 동일한 분량의 과제가 주어졌다. 그리고 문장 낭독 과제는 과제의 특성상 모든 화자들에게 동일한 분량의 과제가 부과되어야 하기 때문에 연령 집단 사이에 차이를 두지 않았다.

하지만 문단 낭독 과제와 괄호 메우기 과제의 경우는 연령을 고려하여 50 대 이상의 조사 대상자들에게 과제의 분량을 줄여서 수행하도록 하였다. 문단 낭독 과제의 경우는 40 대 이하의 조사 대상자들에게는 본 연구팀이 개발한 3 개 문단을 모두 낭독하도록 하였고, 50 대 이상의 조사 대상자들에게는 본 연구팀이 개발한 3 개의 문단 중 한 개 문단(소위 '야옹이 문단')만을 읽도록 하였다.

또, 괄호 메우기의 경우에도 차등을 두어, 40 대 이하의 조사 대상자는 연구팀이 개발한 5 개 주제 영역을 모두 발화하도록 한 반면에 50 대 이상의 조사 대상자는 이 가운데 2 개 영역을 제외하고 3 개 영역(신상, 가족, 상식)만을 수행하도록 하였다.

4.4. 녹음 환경 및 장비

조사원은 녹음실 혹은 조용한 녹음 공간을 확보하여 최대한 조용한 환경에서 자료를 수집하였다. 녹음에는 고성능 디지털 녹음기인 TASCAM DR-07, SONY PCM-D50, SONY PCM-M10, ZOOM H1 등 4종이 활용되었다. 조사원들에게 주로 지급된 녹음기는 TASCAM DR-07이었다. 조사원들은 해당 녹음기를 가지고 조사원 교육을 받은 후에 해당 녹음기로 조사 대상자의 자료를 수집하였다.

조사원들은 녹음 시 녹음기를 삼각대에 설치하여 고정시킨 후 녹음기에 내장된 마이크를 사용하여 녹음하는 방식으로 자료를 수집하였다. 자료는 44.1kHz의 표본 추출률과 16bit 양자화를 통해 디지털화하였고, wav 형식의 파일로 저장하였다.

4.5. 파일 관리

자료 수집 후에 수합되는 자료는 총 4종으로 피험자 동의서, 피험자 정보표, 사례비 영수증, 녹음 파일이다. 녹음 파일을 제외한, 서면으로 작성된 3종의 자료는 각각 별도의 디렉토리를 만들어 보관하였다. 피험자 정보표의 경우는 해당 내용을 모두 스프레드시트로 정리하여 디지털화함으로써 피험자에 대한 검색이나 통계가 가능하게 하였다.

각 조사 대상자의 녹음 파일은 각 조사 대상자별로, 그리고 각 조사 대상자가 수행한 발화 과제별로 분리해 두었다. 앞서 논의한 바와 같이 조사 대상자들은 총 5 개 발화 과제를 수행하였다. 따라서 각 화자는 5 개의 녹음 파일을 생성하게 된다. 녹음 파일의 파일명은 (2)에 보인 원칙에 따라 붙임으로써 파일명만을 보고도 바로 녹음 자료의 주요 특징을 알 수 있도록 하였다.

(2) a. 파일명 부여 방법

(지역)_(연령)_(성별)_(연도별해당조건별조사대상자일련번호)_(전체일련번호)_(수행과제번호)_(대본버전).wav

- 지역: 2 자리 대문자로 표시
  수도권(SG), 경남권(GN), 경북권(GB), 전남권(JN), 전북권(JB), 충남권(CN), 충북권(CB), 강원권(GW), 제주권(JJ)
- 연령: 2 자리 숫자로 표시
  20 대(20), 30 대(30), 40 대(40), 50 대(50), 60 대 이상(60)
- 성별: 1 자리 소문자로 표시
  남성(m), 여성(f)
- 연도별해당조건별조사대상자일련번호: 연도+일련번호
  예) 1610: 2016 년 10 번째 조사 대상자
- 전체일련번호: 일련번호
  예) 040: 40 번째 조사 대상자
- 수행과제번호: 1 자리 대문자로 표시
  A: 괄호 메우기, B: 문장 낭독, C: 문단 낭독, D: 자유 발화, E: 모음 발성
- 대본버전(문장낭독의 경우): V+2 자리 숫자로 표시
  예) V07: 7 번째 버전(즉, 6 차 수정본)

음성 파일과 관련된 전사 파일, 혹은 레이블링 파일의 경우는 파일명의 앞부분을 공유함으로써 음성 파일과 함께 관리될 수 있도록 하였다. 2017 년 2 월 현재, 모든 음성 파일에 대한 전사 파일이 갖추어진 상태다. 낭독 자료의 경우는 해당 대본이 마련되어 있기 때문에 전사에 큰 어려움이 없었지만, (준) 자유 발화 자료의 경우는 전사 자료를 갖추는 데 많은 시간과 노력이 소요된다. 본 과제의 목표는 발화 자료 수집에 있었기 때문에 모든 전사 자료가 과제의 종료와 함께 완비되지는 못하였다. 하지만 본 연구팀은 2016 년 11 월 과제 종료 후에도 KSS DB의 완성도를 높이기 위해 지속적인 작업을 수행하였고, 그 결과 가장 시간이 많이 소요되는 자유 발화 과제에 대한 철자 전사를 완료하였다. 앞으로 괄호 메우기 과제에 대한 철자 전사도 수행할 계획이다. 또, KSS DB는 지속적인 정련 작업을 통해 전사는 물론, 분절음 단위와 어절 단위의 레이블링 파일을 함께 데이터베이스화하는 것을 목표로 한다.

5. 요약 및 결론

이 연구는 2014년부터 2016년까지 총 3년간 전국 9개 권역, 20세 이상 5개 연령층, 남녀 총 3,000명 이상의 발화를 수집하고 이를 데이터베이스화함으로써 한국어의 음성적 특징을 다각적으로 살펴볼 수 있는 의미 있는 자료를 구축하고 이를 데이터베이스화하는 것을 목표로 하였던 한국인 표준 음성 데이터베이스(Korean Standard Speech Database, KSS DB)의 구축 과정을 상세히 소개하는 것을 목적으로 하였다. 연구를 통해 3년간의 총 3,191명의 발화 자료를 수집하여 데이터베이스화한 과정을 상세히 소개하였다.

수행 결과를 요약하면 다음과 같다. 1 차년 전반기에는 국내외 음성 데이터베이스의 현황을 파악하고, 간접 조사 방법으로 수집 가능한 자료를 수집함과 동시에 연구 목표, 즉 한국인의 음성적 특징을 다각적으로 살펴볼 수 있는 음성 데이터베이스를 구축할 수 있도록 데이터베이스를 설계하고 자료 수집 프로토콜을 개발하는 데 총력을 기울였다. 간접 수집 방법을 통해 모두 362 명의 발화를 수집하였는데, 대체로 20 대 화자의 발화 자료가 주를 이루었다. 20 대 화자의 자료는

281 명분으로 전체 자료의 74%에 해당하였다.

1 차년 후반기부터는 직접 수집 방법을 통해 자료를 수집하였다. 연구팀은 전국을 9 개 지역 방언권(수도권, 경남권, 경북권, 전남권, 전북권, 충남권, 충북권, 강원권, 제주권)으로 나누어 해당 지역에서 나고 자라서 현재 해당 지역에 거주하고 있는 사람들을 단일 방언권 화자로 정의한 후에 이들을 주 조사 대상자로 삼았다. 직접 수집 조사 대상자 총 2,829 명 중 단일 방언권 화자는 총 2,615 명이었다.

연구팀은 수도권으로 이주한 방언 화자들의 음성적 특징에도 관심을 두어 단일 방언권 화자 외에도 복합 방언권 화자를 섭외하여 자료를 수집하였다. 수도권 이주 연한에 따라 5 년 이상 15 년 미만 화자와 15 년 이상 화자의 두 그룹으로 나누어 총 205 명의 복합 방언권 발화 자료를 수집하였다.

한편, 단일 방언 혹은 복합 방언 화자의 요건을 갖추지 못한 조사 대상자의 경우는 기타 방언권으로 분류하였다. 기타 방언권에 속하는 조사 대상자는 모두 연속 조사 대상자들이었다. 연속 조사 대상자란 연구 과제가 진행된 3 년 동안 8 회 이상의 녹음을 주기적으로 수행한 조사 대상자들을 말한다. 연속 조사 대상자는 지속적인 녹음 가능성이 가장 중요했기 때문에 일부 화자의 경우 단일 방언 혹은 복합 방언의 조건에 맞지 않는 화자들이 일부 포함되었다. 그 수는 가능한 한 최소화하려고 하였다. 그 결과 기타 방언권 화자는 직접 조사 대상자 총 2,829 명 중 9 명에 불과했다.

인구 통계를 반영하면서도 각 지역, 연령, 성별 조건에 맞는 최소한의 인원수를 확보하기 위하여 1 차년과 3 차년에는 인구 통계 비례를 반영한 조사 대상자의 목표치를 설정하였고, 2 차년에는 균등 배분 방식으로 조사 대상자의 목표치를 설정하였다. 이를 통해 각 지역, 연령, 성별 조건에 맞는 최소 조사 대상자의 수를 확보하려 노력하였으나, 대체로 지역적으로는 강원과 충북권에서 목표보다 적은 조사 대상자가 섭외되었고, 연령적으로는 30 대 화자가 목표에 미치지 못하였다.

전국 단위 자료 수집을 위하여 연구팀의 연구원들 외에도 각 지역에서 조사원을 모집하였다. 모집된 조사원들은 연구팀이 마련한 교육 과정을 이수한 후에 자료 수집에 들어갔다. 본 연구를 위해 연구팀의 연구원 외에 3 년간 총 124 명의 조사원들이 활동하였다.

조사 대상자로부터 한국어의 음성학적인 특징을 잘 살펴볼 수 있는 자료를 효율적으로 수집하기 위해 조사 자료를 설계하였다. 연구팀은 조사 대상자들이 수행할 총 5 가지 발화 과제를 개발하였는데, 이중 2 가지는 낭독 발화 과제(문장 낭독 과제와 문단 낭독 과제)였고 3 가지는 준 자유 발화(모음 발성, 괄호 메우기) 혹은 자유 발화 과제였다.

과제의 제시 순서는 모음 발성, 문단 낭독, 문장 낭독, 괄호 메우기, 자유 발화와 같았는데 과제 수행의 난이도를 고려하여 그 순서를 설계하였다. 또한, 원활한 조사를 위해 조사 대상자의 연령에 따라 과제의 분량을 조절하였다. 40 대 이하의 조사 대상자들은 5 가지 과제 모두 전체 분량을 수행하게 한 반면에 50 대 이상의 조사 대상자들은 문단 낭독에서 2/3, 괄호 메우기에서 2/5 의 분량을 줄여서 수행하게 하였다. 하지만 모음 발성, 문단 낭독, 자유 발화의 경우는 연령에 따른 수행 분량 차이를 두지 않았다.

모든 음성 자료는 44.1kHz의 표본 추출률과 16bit 양자화를 통해 디지털화하였고, wav 형식의 파일로 저장하였다. 파일명은 녹음 자료의 주요 특징을 알 수 있도록 부여하여 관리와 식별의 편의를 도모하고자 하였다.

3 년간의 연구를 통해 구축된 KSS DB는 규모 면에서나 설계 면에서 한국어의 음성적 특징을 알아보기 위한 음성 코퍼스가 갖추어야 할 대표성과 균형성을 고루 갖추었다고 할 수 있다. 하지만 DB의 완성도를 높이기 위해서는 지속적인 노력이 필요하다. 이에 고려대학교 연구팀은 과제 수행 기간 종료 이후에도 자체 재원을 통해 상대적으로 수집 인원이 부족한 지역, 성별, 연령을 보완하기 위해 추가 자료 수집은 물론, 데이터 베이스 정련 작업을 지속적으로 수행하고 있으며, 자유 발화 과제에 대한 철자 전사 작업을 수행하고 있다. 또한, 음성 파일에 대한 분절음 단위와 어절 단위의 레이블링을 진행함으로써 DB를 활용한 대규모 음성 자료를 바탕으로 한 음성학 연구가 다차원적으로 이루어질 수 있는 기초를 마련하고자 한다.

참고문헌

Shin, J., Jang, H., Kang, Y., & Kim, K. (2015). Developing a Korean Standard Speech DB. *Phonetics and Speech Sciences*, 7(1), 139-150. (신지영·장혜진·강연민·김경화 (2015). 한국인 표준 음성 DB 구축. *말소리와 음성과학*, 7(1), 139-150.)

National Institution fo Korean Language (2007). *21st Century Sejiong Project Developing Special Data of Korean Language*. Seoul: National Institution fo Korean Language. (국립국어원 (2007). *21 세기 세종계획 국어 특수자료 구축*. 서울: 국립국어원.)

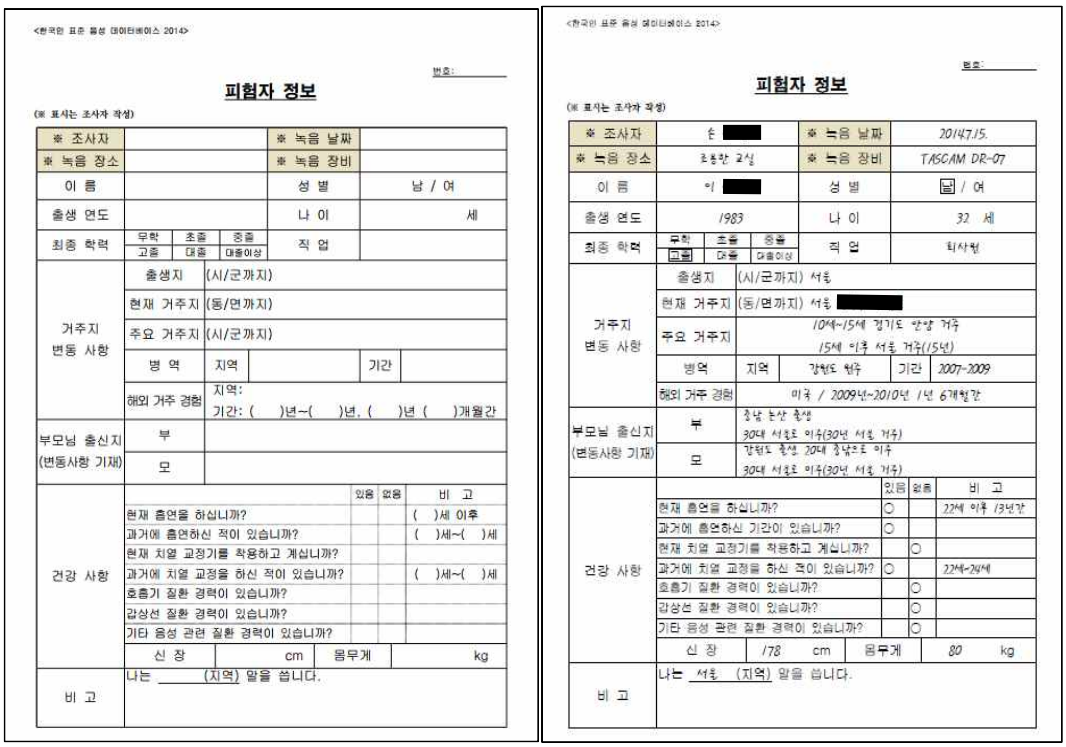
• **신지영 (Shin, Jiyoung)** 교신저자
고려대학교 국어국문학과
서울시 성북구 안암로 145
Tel: 02-3290-1973
Email: shinjy@korea.ac.kr
관심분야: 음성학, 음운론

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

부록 1. KSS DB 전체 직접, 간접 조사 대상자의 지역별, 연령별, 성별 총괄표 (단위: 명)

			20대		30대		40대		50대		60대+		지역별 성별 소계		지역별 소계			
			남	여	남	여	남	여	남	여	남	여	남	여				
직접조사	단일방언권	수도권	64	65	62	83	40	105	37	76	15	32	218	361	579	수도권	단일방언권	직접조사
		경남권	60	71	38	37	24	52	30	50	15	15	167	225	392	경남권		
		경북권	41	70	20	36	40	58	25	48	11	18	137	230	367	경북권		
		전남권	50	53	14	24	15	54	25	36	7	12	111	179	290	전남권		
		전북권	43	67	14	23	11	29	33	26	13	9	114	154	268	전북권		
		충남권	44	47	15	24	27	24	17	22	17	12	120	129	249	충남권		
		충북권	20	24	8	14	10	27	17	27	4	7	59	99	158	충북권		
		강원권	27	49	5	10	8	21	11	11	4	4	55	95	150	강원권		
		제주권	13	49	29	11	9	13	9	13	8	8	68	94	162	제주권		
	소계		362	495	205	262	184	383	204	309	94	117	1,049	1,566	2,615	소계		
	연령별 합계		857		467		567		513		211		2,615		연령별 합계			
	복합방언권	경상	43	27	12	14	7	4	4	7	0	4	66	56	122	경상	복합방언권	
		전라	13	12	6	8	6	11	4	20	2	1	31	52	83	전라		
	소계		56	39	18	22	13	15	8	27	2	5	97	108	205	소계		
	기타		1	1	2	0	1	1	1	2	0	0	5	4	9	기타		
	소계		1	1	2	0	1	1	1	2	0	0	5	4	9	소계		
	성별 계		419	535	225	284	198	399	213	338	96	122	1,151	1,678	2,829	성별 계		
	연령별 계		954		509		597		551		218		2,829			연령별 계		
전체총계	연령별 총계		1,223		543		617		579		229		3,191		연령별 총계			전체총계
	성별 총계		547	676	255	288	198	419	224	355	105	124	1,335	1,862	성별 총계			
간접조사	연령별 소계		269		34		20		28		11		362		연령별 소계			간접조사
	수도권		128	141	30	4	0	20	11	17	9	2	179	184	362	수도권		
			남	여	남	여	남	여	남	여	남	여	남	여	지역별 소계			
			20대		30대		40대		50대		60대		지역별 성별 소계					

부록 2. 피험자 정보표 양식(왼쪽)과 작성 예(오른쪽)



부록 3. 발화 과제 상세 정보

I. 문장 낭독 55 개 문장(최종본 기준)

- 모음 문장과 자음 문장 표에서 첫 번째 칸은 목표 음운을, 두 번째 칸은 문장 번호를, 세 번째 칸은 해당 문장을 각각 의미한다.
- 운율 및 음운 현상 표에서 문장의 첫 번째 칸은 문장 번호를, 두 번째 칸은 해당 문장을 각각 의미한다.
- 문장 낭독은 문장 번호 순서로 진행되었다.

a. 모음 문장

단모음		
ㅏ	54	아들과 병원에 가보니 하필 늑막염이었다.
ㅐ	14	애달픈 개처럼 해 질 녘까지 엄마를 기다렸다.
ㅓ	46	어제는 허리가 아파서 거동이 불편했다.
ㅔ	16	에누리 없이 게와 고등어를 파는 사람들을 헤아려 보았다.
ㅗ	45	오전에 호루라기를 챙겨서 고사장으로 갔다.
ㅜ	38	우리는 후식으로 과자를 먹으며 구름을 보았다.
ㅡ	44	으름장을 놓아서 그를 흐느끼게 만들었다.
ㅣ	7	이동할 때를 기다리며 히히덕거렸다.
이중모음		
ㅑ	9	야생화 향기를 맡아보더니 고개를 약간 갸웃거렸다.
ㅒ	20	애는 애기꾼이라 개보다 애기를 잘해.
ㅕ	17	여러 가지 현악기가 걸이 곱다.
ㅖ	47	예금 상품의 혜택을 계산해보았다.
ㅘ	15	화전된 과거의 일이 화를 불러왔다.
ㅙ	31	왜가리가 괘씸하게도 횃대를 부러뜨렸다.
ㅚ	37	외삼촌은 금융업 분야와의 회식을 괴로워했다.
ㅛ	32	요즘 교사들 사이에서 효도 관광이 유행이다.
ㅝ	36	워낙 권력을 좋아해서 훨씬 더 회장 자리에 집착했다.
ㅞ	29	웨딩드레스에 대해 궤변을 늘어놓으며 훼방을 놓았다.
ㅟ	42	위대한 귀농인을 만나기 위해서 휘파람을 불며 옷 입고 나갔다.
ㅠ	1	유도부가 휴가지에서 귤을 사 왔다.
ㅢ	41	의사는 축의금과 함께 희소식을 전했다.

b. 자음 문장

ㄱ	27	가평의 고추밭에서 기적적으로 구출되었다.
ㄲ	51	까치는 꾸물거리다가 구멍 사이에 끼었다.
ㄴ	23	나무에 널린 니트들이 누구의 것인지 궁금했다.
ㄷ	2	다리 아래 사는 두더지는 땅을 다딜 수 없었다.
ㄸ	3	따가운 자외선에 뚜껑이 검은 빛을 띠었다.
ㄹ	50	라면을 먹으면서 솜이불 위에서 루마니아 축구 리그를 보았다.
ㅁ	39	마침내 무서운 미로에서 빠져나왔다.
ㅂ	34	바다에서 부드러운 바람을 맞으며 비빔밥을 먹었다.
ㅃ	33	빠르게 뛰다가 뿌리에 걸려서 발목을 삐었다.
ㅅ	52	사다리 위에서 수다를 떨다 보니 시간이 다 되었다.
ㅆ	49	싸다고 하는 씨감자로 죽을 쑤다 말았다.
ㅈ	48	자전거 가게에서 주민 회의가 지금 진행 중이다.
ㅉ	5	짜게 끓인 찌개를 쭈그려 앉아서 먹었다.
ㅊ	11	차선을 넘나들며 치열하게 추격전을 벌였다.
ㅋ	24	카메라와 쿠키를 들고 있는 남자가 가장 키가 크다.
ㅌ	43	타조는 투명한 유리에서 칼날과 티끌을 발견했다.
ㅍ	22	파란 눈과 하얀 피부 덕분에 첫인상이 푸근해보였다.
ㅎ	53	하늘이는 후미진 골목 끝에서 히죽거렸다.

c. 운율 및 음운 현상

4	아기 옷을 벗기고 입히면서 빗으로 머리도 빗겨 주었다.
6	밝은 빛으로 곱하기 공부를 하다가 급하게 책 한 권을 읽었다.
8	큰 빚을 져서 맛있는 찜닭도 못 먹고 밤낮으로 뛰어다녔다.
10	사기그릇 가게로부터 사기를 당했다는 걸 알고 그들은 사기가 떨어졌다.
12	서른여덟의 김유신은 권력을 이용해 불법으로 생산라인을 개조하였다.
13	소주와 김밥과 통닭을 즐겨 먹은 까닭에 수일 내에 돼지가 될 것 같다.
18	원룸에 사니까 절약할 필요가 없어서 방을 밝게 한다.
19	일요일날에는 안암 1 동에서 2 동으로 이동하는 것도 은근히 일이다.
21	저 병에 든 약을 마시면 병에 차도가 있을 것이다.
25	음운론을 가르치시던 담임선생님은 흙에서 넓죽한 고구마를 캐셨다.
26	서울역에서 김연아를 보고 그 찬란한 모습에 넋이 나갔다.
28	산기슭에 있는 장미꽃으로 장식을 하려다가 가시의 끝을 만졌다.
30	늦여름이나 가을날에 상견례를 하려고 온라인으로 식당을 예약했다.
35	넓게 지어진 연륙교가 효과적으로 제 몫을 다 해내고 있다.
40	이 모 씨의 이모가 마침내 고소 절차를 밟게 되었다.
55	의견서 작성 시 띄어쓰기의 원칙에 주의해 주십시오.

II. 문단 낭독 3 개 문단

- 40 대 이하: 아래 3 개 문단 모두 수행
- 50 대 이상: 아래 3 번 야옹이 문단만 수행

1. 미영이 나연이 문단

> 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.

2. 라면 문단

> 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.

3. 야옹이 문단

> 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.

III. 괄호 메우기: 5 개 영역

- 40 대 이하: 5 개 영역 모두 수행
- 50 대 이상: 신상, 가족/친구, 상식 3 개 영역만 수행

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...).
1 년은 _____일이고,
1 주일은 ____일이고, 하루는 ____시간입니다.