Design and Construction of Korean-Spoken English Corpus (K-SEC)*

Seok-Chae Rhee1, Sook-Hyang Lee2, Seok-Keun Kang2, Young-Ju Lee3

1Department of English Language and Literature, Yonsei University
2English Department, Wonkwang University
3Department of Electrical, Electronic and Information Engineering, Wonkwang University

scrhee@yonsei.ac.kr, {shlee, skkang, yjlee}@wonkwang.ac.kr

Abstract
K-SEC(Korean-Spoken English Corpus) is a kind of speech database that is under construction by the authors of this paper. This article discusses the needs of the K-SEC from various academic disciplines and industrial circles, and it introduces the characteristics of the K-SEC design, its catalogues and contents of the recorded database, exemplifying what are being considered from both Korean and English languages' phonetics and phonologies. The K-SEC can be marked as a beginning of a parallel speech corpus, and it is suggested that a similar corpus should be enlarged for the future advancements of the experimental phonetics and the speech information technology.

1. Purpose
Korean-Spoken English Corpus (K-SEC) is a digitally recorded database of English speech sounds spoken by Koreans, which can be reproduced through computer. The purpose of this study is to record a large amount of English sounds spoken by Koreans and to make a data base of it, which can provide basic materials to study in experimental phonetics, phonology, English education and speech information technology. In this paper, we will show how the corpus is designed and made.

2. The background and necessity of K-SEC
In light of experimental phonetics, it is really meaningful to study how the phonological and phonetic features of Korean interfere with Korean speakers' pronunciation of English. Unfortunately, however, there exists no speech corpus which experimental phoneticians can share and utilize for their study. Some researchers record a small number of speech sounds for their own study, but this may lead to a lack of objectivity in the results. In addition, it is very uneconomical to record sounds whenever necessary. Considering these facts, it is suitable to develop a corpus containing a large number of English sounds spoken by Koreans, which can be used as materials for experimental phonetics.

Today the English education system in Korea is shifting its focus from reading to speaking. In speaking, it is important to correctly pronounce English words. In Korea, however, English pronunciation is taught without considering what English sounds are problematic with Korean speakers. Thus there is room for doubt that the current pronunciation education will produce satisfactory results. In this study, we will construct a large corpus which can help to effectively teach English pronunciation by showing clearly and objectively what English sounds Korean speakers have difficulty in pronouncing. Subjects will be divided according to region, sex, age and degree of English education. The corpus to be developed in this study will provide materials for sociolinguistic and psycholinguistic studies. In addition, the technology and results of this study also can be used as important materials for Speech Information Technology, SIIT. Currently around the world, there is an increasing trend to realize the conversation between computer and human beings by using speech sounds, and the outlook for the technology is looking up in industry. A Korean-spoken English corpus is essential to the development of speech perception technology and English learning software program in the field of SIIT(Speech Technology in Language Learning).

3. Design
Before recording, we decided a list of test words as well as the number and types of speakers. In the selection of the test words, the following factors were considered:
- Record read speech.
- Record lexical speech sounds(isolated words) and connected speech sounds (sentences and stories)
- Record standard speech sounds.
- For lexical speech sounds, choose PBW(Phoneme Balanced Words).
- Classify lexical speech sounds according to subject's age (elementary school students, middle/high school students, adults)
- In the case of connected speech sounds, use the same test materials regardless of subject's age.
- Consider the segmental and suprasegmental elements of both lexical and connected speech sounds.
- Design the DB such that it can be connected with overseas speech DB.

The speakers of K-SEC were divided into 3 groups by age (5th and 6th graders/high school students/adults), 7 groups by region(Seoul/Gyeonggi/Gangwon/Chungcheong/Jeolla/Gyeongsang/Jeju), and 2 groups by sex (male/female), and each group consisted of eight speakers. In total, 336 Korean speakers participated in the recording of K-SEC.

(3/age x 7/region x 2/sex = 42 groups) x 8/group = 336

*This paper was supported by Korea Research Foundation Grant (KRF-2001-042-A00035).
Designed this way, the test will clearly show the sociolinguistic characteristics of pronunciation, the variation of pronunciation depending on the degree of English education, and the differences between male and female.

For comparison, on the other hand, native speakers of English were recorded as follows: 2/age x 3/region (North American English, British English, Australian English) x 2/sex x 2/group = 24 subjects.

The test list was made up of words, sentences and stories which can represent the segmental and suprasegmental features of English well. The sentences and stories contained words or sequences of words with the following characteristics of English phonology, which are generally assumed to show the interference of Korean:
- the distinction between labial stops and labio-dental fricatives
- the distinction between voiced and voiceless obstruents
- the relation of voicing of obstruents with the length of preceding vowels
- the aspects of flapping
- the degree of aspiration of stops after [s]
- the degree of aspiration of stops in different positions within words
- the degree of release of plosives in different positions within words
- the distinction between alveolar fricatives and palato-alveolar fricatives/affricates
- the degree of partial devoicing of obstruents
- the aspects of sandhi
- labialization
- sequences of consonants and nasals
- pronunciation of intervocalic voiceless stops
- the distinction between [n], [r], and [l]
- the aspects of schwa
- rhythm
- intonation
- pause

The K-SEC is a data base of English pronunciations spoken by Koreans. In order to compare the sounds of native language L1 with those of foreign language L2, basic Korean speech sounds were also recorded.

Based on the principles and factors mentioned above, the test list was made up of the following six sets:
- Set#1: Korean isolated words (real words and nonce words, some of which were put in a carrier phrase)
- Set#2: Korean stories
- Set#3: English isolated words to test English consonants and vowels(some of them were put in a carrier phrase)
- Set#4: English isolated words (Different word sets were used for elementary school students, middle/high school students and adults respectively.)
- Set#5: English sentences
- Set#6: English stories

3.1 Set#1

Set#1 was designed to compare Korean consonants and vowels with English consonants and vowels. Eighteen Korean consonants which can occur word-initially were put in the frame [Cada], seven monophthongs were put in the frame [hVd], and tense and lax alveolar fricatives were put before high front vowels. All the test words were repeated twice and randomized by using a computer software. In addition, consonants were put in the frame [Cada] in order to see the characteristics of each consonant, and [가라], [겨라] and [타라] were added to the list in order to observe liquid in intervocalic position. The resulting (real and nonce) words were repeated twice, randomized and put in a carrier phrase. Set#1 was prepared as described above in order to compare the consonants and vowels of Korean with those of English. The results of recording Set#1 can be used for the contrastive study of Korean and English speech sounds as well as for the experimental study of Korean itself.

Set#1: Korean isolated word list (For limited space, only some of the words are provided here.)

바다, 호타, 까다, 해다, 나다, 카다, 라다, 빗다, 가다, 이것도 타라다, 이것도 빗이다

3.2 Set#2

Set#2, which contained a Korean story, was designed to compare Korean consonants and vowels with English counterparts in connected speech, especially the differences in prosody which appear when the same speaker reads Korean and English. The story 'The Wind and the Sun', which was previously used for a small speech database, was chosen for this experiment in order to make a large database containing more speakers. The English translation of the story is presented in Set#6.

Set#2: Korean story

마람과 땅님이 서로 힘이 더 세다고 나두고 있음을 봤다. 한나그네가 따뜻한 외투를 입고 걸어왔습니다. 그들은 누군가도 나그네의 외투를 먼저 벗기는 이가 힘이 더 세다고 하기로 결정했습니다. 복종은 힘들었으나 불편한 소리 내가 난나그네는 외투를 단단히 아꼈습니다. 이 배달님이 떠زان한 양말을 가만히 내려_pes나 나그네는 외투를물은 빛었습니다. 이하하여 복종은 헹님이 힘이 더 세다고 인정하지 않을 수 없었습니다.

3.3 Set#3

Set#3 was designed to observe English consonants and vowels in the same environments. First, in order to observe vowels in isolated words, ten English vowels were put in the frame [hVd]. The words were repeated twice in a quasi random order. Second, in order to observe consonants, minimal pairs taken from Ladefoged(2001) were used as test materials. The words were repeated twice in a quasi random order and put in a carrier phrase. The sentences in Set#3 were selected in order to see the differences of characteristics of each consonant and vowel depending on English contrastive stress. Here the words to test vowels were taken from Peterson & Barney(1952), and the sentences with contrastive stress were taken from Turk & White(1999).

Set#3: English consonants and vowels in the same environments & sentences

heard, had, heed, hod, heard, hud, hid, hood, had, who'd, Say kye, too; Say by, too; Say chime, too; Say guy, too;
Set#4 was divided into three categories by speaker's age: elementary school students, middle/high school students and adults. This was mainly because, if the same set of words were used for all the subjects, elementary school students would have difficulty in reading some words. In fact, this problem occurred in a pre-test recording. We classified the test words into three levels based on the elementary/middle/high school curriculums announced by the Ministry of Education (No. 1997-15) and Comprehensive Curriculum of Basic Skills published by American Education Publishing. We tried to use as many phoneme balanced words in the list as possible. Twenty four consonants and twenty one vowels (including diphthongs, triphthongs and rhotacized vowels) were used in the test words. The purpose of this test was to see and analyze how Korean speakers pronounce i) consonants before vowels, ii) consonants after vowels, iii) consonants between unstressed vowels and stressed vowels, iv) consonants between stressed vowels and unstressed vowels (i.e., CV, VC, VCV(+stress), V(+stress)CV). We first tried to choose the data from the words listed for elementary school students. If we could not find a relevant word from the list, we tried to choose the word from the list for middle/high school students, and if we still could not find the word, then we used the word list for adults. In the case of the speech lists for middle/high school students and adults, supplementary words were added, which were taken from Extended Hughes and Trudgill's Word List(EWL, 2002) in consideration of its comparison with an American speech database and a Japanese-spoken English database.

In addition, words showing the pronunciation aspects of word-initial clusters were added. The list also included words which show the insertion of stop between a nasal and a fricative (e.g., print/prince), words which show the relation of stress with the number of syllables and vowel length (e.g., speed/speedy/speedily), and words which show the pronunciation of glottal stop (e.g., button). We collected 435 words for elementary school students, 956 words (including the 435 words for elementary school students) for middle/high school students, and 1125 words (including the 956 words for middle/high school students) for adult speakers, which were all randomized.

Set#4: adult

hobby, keep, heat, jam, dig, soap, mouse, tobacco, map, dining, pork, cook, dark, quiz, so, rethink, follow

Set#5, which contained a list of English sentences, was designed to see i) Korean speakers' intonation and rhythmic patterns in English connected speech (e.g., declarative sentences, Wh-questions, Yes-no questions, tag questions, alternative questions, emphatic expressions, exclamatory sentences, etc.), ii) the degree of sandhi between words, iii) the errors which Korean speakers are apt to make in pronunciation of segments (e.g., incorrect voicing between voiced sounds, incorrect nasalization, etc.), iv) how the phenomena observed in Set#3 and Set#4 are realized in sentences. Set#5 was made up of basic words so that it could be used for all the three different levels of subjects: i.e., elementary school students, middle/high school students and adults.

Set#5: English sentence reading

Miss Henry drank a cup of coffee.
What are you looking for?
Put your toys away right now.

3.6 Set#6

Set#6 was designed to observe various prosodic characteristics in a story.

Set#6: English story reading

The Wind and the Sun were discussing who was the stronger of the two. Since they could not agree, they decided to test their strength. They saw a traveler walking on the road and the Sun said, "Whoever can make him take off his coat will be judged the strongest. You go first." So the Wind blew as hard as he could, but no matter what he did, he couldn't make the traveler take off his coat. Instead, the traveler wrapped his coat more tightly around himself. The Wind finally gave up. Now it was the Sun's turn. The Sun came out and began to shine on the traveler. At first the traveler unbuttoned his coat. Then, as the air grew warmer and warmer, the traveler took off his coat altogether. The Sun turned to the Wind and said, "You see, persuasion is better than force."

4. Recording

4.1 Recording Equipment and Facility

Recording was done in the sound-treated booths at the following locations: Yonsei University in Seoul for Seoul and Gyeonggi Province speakers, Yonsei University in Wonju for Gangwon Province speakers, Wonkwang University for Jeonla Province speakers, Korea Maritime University for Gyeongsang Province speakers, and Jeju National University for Jeju Province speakers. The following equipments were used for recording at Wonkwang University:

- digital recorder: TASCAM DA-20MKII
- audio mixer: Behringer MXB1002
- headset microphone: Sennheiser HMD25-1
- DAT tapes: SONY 10DT-120RA

For recording at the other locations, a portable DAT (TASCAM DA-P1) and a SHURE SM10A microphone were used.

4.2 Preparation of speakers for recording
• The speech list was given to speakers 3 or 4 days before recording so that they could practice it in advance. Speakers were allowed to mark information such as phonetic symbols and pause on the list whenever necessary for practice. It, however, was informed that they could not use any markings in real recording.
• Fifth and sixth graders, high school freshmen and sophomores, and college students and graduate students were chosen as subjects for recording.
• Those who had lived in English-speaking countries more than one year were not included as subjects.
• Before recording, the information of recording date, recording engineer, used equipment, speaker and recording environment was recorded, which was later included in the corpus as a database.

4.3 Preparation before recording
• Three people monitored the recording as a team. (One checked pronunciation, another checked volume and equipment, and the other supervised the whole process of recording.)
• Before recording, speakers were asked to fill out the 'Information of Speaker' form. Speakers were also asked to read the information that they provided for recording.
• Speakers were asked to give detailed information of their learning of English.
• Speakers were asked to read as naturally as possible.

4.4 Main Recording
• Before recording Set#3, speakers' pronunciation ability of English was checked using the words in Set#3. In so doing, volume was also controlled.
• Speakers were told not to use a rising intonation at the end of each word. Speakers were also asked to pause between words or sentences.
• When a word or sentence was mispronounced, only the relevant part was read again. In the case of Korean and English stories, only the mispronounced sentences were re-recorded. (When a speaker realized his or her mistake and read the relevant word or sentence again, the speaker was not required to re-read it later. The re-read word or sentence, however, was recorded in the 'error list sheet'.)
• Either when a speaker missed or mispronounced a word or sentence or when there was a sudden change in the speaker's voice, the relevant word or sentence was re-recorded. Re-recorded words or sentences were also put in the 'Re-recording List Sheet'.
• In word recording, when inflectional suffixes such as plural suffix 's' and past tense suffix 'd' were not pronounced, the relevant words were re-recorded. In the case of sentence recording, however, such errors were ignored.
• Speakers were asked to take a rest at the end of each set. In the case of Set#4, speakers were required to take a break after every one-third of the list. Speakers, however, were allowed to take a rest whenever they wanted to.
• After recording, speakers were asked to read and sign the 'Confirmation and Receipt Form' (attached form #6) and also to fill out a separate 'List of Speakers'.

4.5 Processing Plan after Recording
The recorded materials will be digitalized and made into systematic files with names for effective search.
• Speech sounds will be selectively segmented and labelled under the support of analysis equipment by the SiTEC at Wonkwang University.
• Transcription will be selectively done, and both orthographic and phonetic levels will be considered for it. This is to make the results of the study useful to phoneticians, phonologists and English pronunciation educators.
• Prosodic transcription will also be done according to the labelling system of ToBI in order to embody Korean speakers' errors in pronouncing English intonation objectively and systematically.
• To sum up, the corpus will include:
  - signals (speech signals and also some laryngeal signals)
  - characteristics (the characteristics of speakers and recording)
  - various marks (marks which represent pitch range and the beginning of vowels)
  - notes (transcription marks)

5. Conclusion
In this paper, we have described i) the background and necessity of construction of K-SEC, and ii) the characteristics of its design and the construction process of it. Now the design of the K-SEC has been completed, and recording has almost been finished. After recording is completed, the results will be filed according to each speech unit (word, sentence, paragraph) under the support of the SiTEC, and representative speeches will be selectively transcribed. Once the corpus is completed, it will be distributed as a CD by SiTEC. The K-SEC can be marked as a beginning of a parallel speech corpus, and it is suggested that a similar corpus should be enlarged for the future advancements of the experimental phonetics and the speech information technology.

6. References