Korean Children's Spoken English Corpus and an Analysis of its Pronunciation Variability

Hyejin Hong[1], Sunhee Kim[2], Minhwa Chung[1]
[1]Department of Linguistics, Seoul National University, Seoul 151-742, Republic of Korea
[2]Center for Humanities and Information, Seoul National University, Seoul 151-742, Republic of Korea
E-mail: {souble1, sunhkim, mchung}@snu.ac.kr

Abstract

This paper introduces a corpus of Korean-accented English speech produced by children (the Korean Children's Spoken English Corpus: the KC-SEC), which is constructed by Seoul National University. The KC-SEC was developed in support of research and development of CALL systems for Korean learners of English, especially for elementary school learners. It consists of read-speech produced by 96 Korean learners aged from 9 to 12. Overall corpus size is 11,937 sentences, which amount to about 16 hours of speech. Furthermore, a statistical analysis of pronunciation variability appearing in the corpus is performed in order to investigate the characteristics of the Korean children's spoken English. The realized phonemes (hypothesis) are extracted through time-based phoneme alignment, and are compared to the targeted phonemes (reference). The results of the analysis show that: i) the pronunciation variations found frequently in Korean children's speech are devoicing and changing of articulation place or/and manner; and ii) they largely correspond to those of general Korean learners' speech presented in previous studies, despite some differences.

Keywords: Korean-accented English corpus, children's speech, pronunciation variability

1. Introduction

Due to recent advances in speech and language technology, there is a growing interest in incorporating the techniques into computer-assisted language learning (CALL) systems (Eskenazi, 2009). When deploying automatic speech recognition (ASR) technique in CALL for non-native learners, it should be guaranteed that the speech produced by non-native learners is properly recognized and evaluated in order to give learners appropriate feedback on their performance. However, speech recognition performance of non-native speech degrades severely compared to native speech recognition performance due to variations in non-native speech (Benzeghiba et al., 2007).

Dealing with non-native's pronunciation variability requires an appropriate speech corpus and analysis of pronunciation variations based on the corpus. There are several speech corpora which specialize on non-native speech (Gruhn et al., 2011). However, the number of non-native speech corpora is lower than that of native speech corpora.

This paper introduces a corpus specializing in Korean-accented English speech, the Korean Children's Spoken English Corpus (KC-SEC)[1]. The KC-SEC was constructed by Seoul National University in support of research and development of CALL systems for Korean learners of English, especially for elementary school learners. The statistical analysis of the English pronunciation variability appearing in the KC-SEC corpus is also shown to describe the characteristics of the corpus. The KC-SEC corpus and the analysis results are beneficial to improvement of ASR performance and research on language learning.

[1] The KC-SEC corpus is available via Speech Information Technology & Industry Promotion Center (sitec@wku.ac.kr or dlchoi@wku.ac.kr) at Wonkwang University in Korea.

2. KC-SEC Corpus

Although some speech data collections including the Speech Accent Archive (Weinberger, n.d.) and the corpus used for Witt (1999) contain Korean-accented English speech, so far, the only publicly available resource which specialized in Korean-accented spoken English is the Korean-Spoken English Corpus (K-SEC) (Rhee et al., 2004). The K-SEC consists of isolated words and a small number of sentences uttered by three groups of speakers: children (5th and 6th graders), high school students, and adults.

With a strong need of Korean-accented English speech corpus consisting of a large number of sentences uttered by children for developing a dialog-based CALL system, we constructed a new read-speech corpus called the KC-SEC (Kim et al., 2010). As shown in Table 1, the KC-SEC includes read-speech produced by a total of 96 Korean learners of English, aged from 9 to 12[2]. The speakers were recruited so that age, gender and English education experience are equally distributed.

Age	9~12 years (mean = 10.76, SD = 0.99)
Gender	Male: 45 Female: 51
English Education	0.5~9.5 years (mean = 3.53, SD = 1.85)

Table 1: Speaker information of the KC-SEC.

The KC-SEC corpus contains 704 unique sentences from various English textbooks for Korean children as well as all the 36 sentences used for the K-SEC corpus for

[2] A total of 99 Korean learners had participated in the recording. However, speakers (2 males and 1 female) were excluded because of recording errors.

comparison purpose. The total vocabulary size is 858 words. Each sentence consists of the words from 2 to 14, with an average of 6.09 words (SD = 1.96). Most sentences (96.82%) do not exceed 10 words as illustrated in Figure 1.

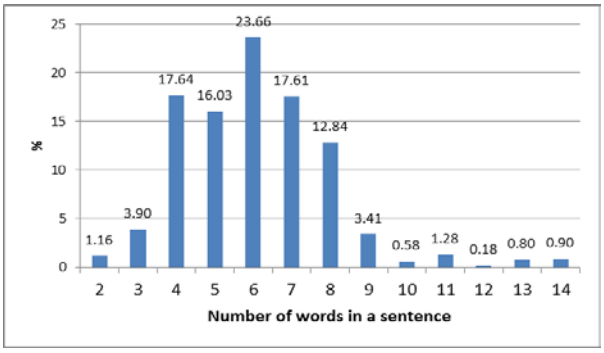

Figure 1: Distribution of number of words in a sentence.

Examples of sentences are as follows.

(1) Yes, can you get me a cup of tea?
(2) That's a very nice name.
(3) Let me draw you a map.
(4) I'll go to another shop to buy one then.
(5) What kind of fish can I catch here?

The recording was conducted in relatively silent but not sound-proof rooms to mimic the real usage setting[3]. All speakers were requested to read the sentences presented on a screen as shown in Figure 2. A sentence list consisting of 60 common sentences and 104 variable sentences plus 36 K-SEC sentences was given to each speaker. The speakers read the common sentences first and then proceed with the remaining sentences during a 1-hour recording session. The speech was recorded in Windows PCM format with 16 kHz sampling rate and 16 bit quantization, using a headset (Sennheiser PC 151) and an external sound card (Creative Sound Blaster X-Fi Go!).

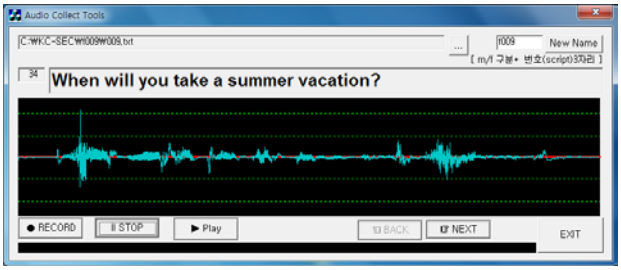

Figure 2: Screenshot of the recording environment.

Depending on the speaker's English proficiency level, 50 to 200 sentences were read by each speaker. As a result, read-speech data of about 16 hours, consisting of 11,937 sentences (72,644 words), were collected.

[3] Despite a slight level of noise, the corpus is good enough to be used for phonetic analysis.

Phoneme	IPA	Number of occurrences	Phoneme	IPA	Number of occurrences
AA	ɑ	4,652 (2.12%)	L	l	7,581 (3.46%)
AE	æ	6,369 (2.90%)	M	m	6,309 (2.88%)
AH	ʌ	15,407 (7.03%)	N	n	14,338 (6.54%)
AO	ɔ	3,175 (1.45%)	NG	ŋ	1,852 (0.84%)
AW	aʊ	2,460 (1.12%)	OW	oʊ	3,330 (1.52%)
AY	aɪ	7,997 (3.65%)	OY	ɔɪ	263 (0.12%)
B	b	4,405 (2.01%)	P	p	3,917 (1.79%)
CH	ʧ	1,574 (0.72%)	R	r	10,535 (4.80%)
D	d	7,137 (3.25%)	S	s	8,658 (3.95%)
DH	ð	6,895 (3.14%)	SH	ʃ	1,237 (0.56%)
EH	ɛ	6,779 (3.09%)	T	t	13,715 (6.25%)
ER	ɝ	5,045 (2.30%)	TH	θ	1,440 (0.66%)
EY	eɪ	4,061 (1.85%)	UH	ʊ	2,180 (0.99%)
F	f	3,248 (1.48%)	UW	u	7,386 (3.37%)
G	g	3,233 (1.47%)	V	v	4,864 (2.22%)
HH	h	4,138 (1.89%)	W	w	5,591 (2.55%)
IH	ɪ	11,418 (5.21%)	Y	j	4,577 (2.09%)
IY	i	7,874 (3.59%)	Z	z	7,095 (3.24%)
JH	ʤ	1,441 (0.66%)	ZH	ʒ	18 (0.01%)
K	k	7,096 (3.24%)			

Table 2: Phoneme distribution in the KC-SEC: A total of 219,290 phonemes include 130,894 consonants and 88,396 vowels.

Table 2 shows each phoneme's distribution in the KC-SEC corpus, when each word is converted into its canonical pronunciation by using the CMU pronouncing dictionary (Weide, 2008) and its phoneme set.

3. Pronunciation Variability of the KC-SEC

3.1 Analysis Method

To describe the characteristics of the KC-SEC, a statistical analysis of the English pronunciation variability is performed. For an analysis of pronunciation variability, both knowledge-based and data-driven approaches can be used (Strik and Cucchiarini, 1999). In a knowledge-based approach, pronunciation variability is analyzed on the basis of comparison between phonemic and phonetic systems of the source language (L1) and the target

language (L2). Assuming that all learners show the same patterns of pronunciation, the variations obtained from this approach can elude some variations which exist in real speech corpus. Furthermore, some variations which are not found in the corpus can appear (Cucchiarini et al., 2011; Strik and Cucchiarini, 1999). On the other hand, in a data-driven approach, pronunciation variations are obtained directly from transcriptions of speech. For a data-driven approach, transcriptions can be produced either manually or automatically. Obtaining manual transcriptions by phoneticians is time- and labor-consuming, and there are problems of disagreement and inconsistency. For an analysis of English pronunciation variability appearing in the KC-SEC corpus, in this paper we use an automatic time-based phoneme alignment method (Le and Besacier, 2005), which uses automatically obtained transcriptions.

The analysis is performed in five steps: (1) the forced-alignment procedure for generating the reference representing the canonical pronunciation, (2) the phoneme recognition procedure for generating the hypothesis indicating the realized pronunciation, (3) the time-based phoneme alignment of the reference and the hypothesis, (4) the generation of a phoneme confusion matrix, (5) an analysis of the phoneme confusion matrix. This method assumes that the misrecognized phonemes correspond to the variations of the reference phonemes, which leads to account for the pronunciation variation of Korean-accented English speech.

A forced-alignment is performed on the KC-SEC data to generate reference phoneme sequences using the CMU pronouncing dictionary version 0.7a and a native speaker's English acoustic model. The acoustic model is constructed according to the process described in Vertanen (2006) using HTK v.3.4 (Young et al., 2006). The number of Gaussians is increased up to 256 for 39 phonemes. Then, we obtain hypothesis phoneme sequences by phoneme recognition using the same acoustic model.

In the following step, the reference and hypothesis phoneme sequences are aligned on the time scale. Figure 3 illustrates an example of the time-based phoneme alignment for the word 'money'. From this example, we can obtain pronunciation variations such as AH→AA and IY→EY.

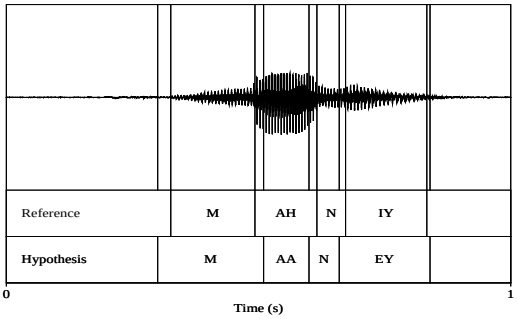

Figure 3: Time-based phoneme alignment for the word 'money'.

Phoneme confusion matrix is generated between the reference and hypothesis phonemes, and the pronunciation variability is analyzed based on it.

3.2 Pronunciation Variability

Pronunciation variations are selected as major, if the rate of mismatch between targeted and realized phonemes is higher than 15.18% and 19.88% for consonants and vowels, respectively. These numbers are determined by 99 percentile of the mismatch rate found in the phoneme confusion matrix. As a result, 7 consonantal variations and 3 vocalic variations are selected as major variations as shown in Table 3.

Reference phoneme	Hypothesis phoneme	%	Oh et al. (2007)
Z	S	33.54	
DH	**D**	**19.62**	Knowledge
N	NG	17.78	
G	K	15.88	
ZH	**JH**	**15.79**	Knowledge
ZH	Z	15.79	
CH	**T**	**15.31**	Data
EH	**AE**	**33.75**	knowledge, data
IH	**IY**	**27.93**	knowledge, data
IH	EY	20.87	

Table 3: The major pronunciation variations of Korean children's spoken English, where the variations also found to be major in Oh et al. (2007) are marked in shaded cells.

Major pronunciation variations of consonants are devoicing (Z→S, G→K), changing of articulation place (N→NG, ZH→Z), changing of articulation manner (ZH→JH) and changing of place and manner (DH→D, CH→T). In the case of vowels, lowering (EH→AE), tensing (IH→IY) and diphthongization (IH→EY) are observed. Some pronunciation variations are related to the English phonemes which do not exist in Korean such as DH, ZH, and IH. Due to the lack of contrast between voiced and unvoiced sounds in Korean, Z and G seem to be problematic for Korean children. English phonemes, CH and EH, which are different in terms of place of articulation, match similar Korean phonemes in most cases. In sum, pronunciation variations found in this study can be explained with lack of L2 phonemes in L1 and difference between L1 and L2 phonemes in terms of place and manner of articulation.

The five variations marked in shaded cells are also found to be major in the previous research on Korean-accented speech corpus (Oh et al., 2007), in which both knowledge-based and data-driven approaches were adopted. The concurrent variations can cover pronunciation variations either from the knowledge-based approach (4 variations) and the data-driven approach (3 variations) of the previous research.

There are unmatched pronunciation variations between

Oh et al (2007) and our results. The variations Z→S, N→NG, G→K, ZH→Z and IH→EY found by us are not found in the previous study, while the variations such as F→P, R→L and AA→AO indicated by the previous study are rarely observed in our corpus. From these comparisons, it can be concluded that, overall, the analysis of the pronunciation variability using the KC-SEC correlates with that of the previous study despite some differences. Further research is needed to investigate whether the common pronunciation variations are L1-related, and the different variations found in this study are age-related or corpus-dependent.

Note that even American English read speech, the TIMIT corpus, shows highly diverse variations (Kim et al., 2011). However, the patterns of pronunciation variations are different. While the pronunciation variations like Z→S and ZH→JH occur with comparatively high frequency ('Z→S' 11.14%, 'ZH→JH' 15.00%), CH→T, EH→AE and IH→EY remain below 0.5%.

4. Conclusion

This paper describes a corpus of Korean-accented English speech produced by children, the KC-SEC, and presents an analysis of pronunciation variability in the corpus. The analysis is based on the time-based phoneme alignment, which compares the targeted phonemes (reference) to the realized phonemes (hypothesis). The statistical analysis using the KC-SEC corpus shows pronunciation variations found frequently in speech of Korean children. Our analysis also shows that the method used in the current study is promising in determining major pronunciation variations.

Based on this initial information on pronunciation variability of Korean children's speech, the pronunciation model of an ASR module of CALL system targeting Korean elementary school learners can be improved. In foreign language learning, this information can be used for setting learning objectives.

5. Acknowledgements

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Knowledge Economy (MKE).

6. References

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49, pp. 763—786.

Cucchiarini, C., van den Heuvel, H., Sanders, E., Strik, H. (2011). Error selection for ASR-based English pronunciation training in 'My Pronunciation Coach'. *Proc. of INTERSPEECH 2011*. pp. 1165—1168.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, 51, pp. 832—844.

Gruhn, R.E., Minker, W., Nakamura, S. (2011). *Statistical Pronunciation Modeling for Non-Native Speech Processing*. Berlin: Springer.

Kim, J., Hong, H., Oh, S., Lee, K., Chung, M. (2010). Design and construction of Korean Children-Spoken English Corpus (KC-SEC). *Proc. of Korean Society of Speech Sciences 2010 Fall Conference*. pp. 120—121 [in Korean].

Kim, S., Lee, K., Chung, M. (2011). A corpus-based study of English pronunciation variations. *Proc. of INTERSPEECH 2011*. pp. 1893—1896.

Le, V.-B., Besacier, L. (2005). First steps in fast acoustic modeling for a new target language: Application to Vietnamese. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*. pp. 821—824.

Oh, Y.R., Yoon, J.S., Kim, H.K. (2007). Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. *Speech Communication*, 49, pp.59—70.

Rhee, S.-C., Lee, S.-H., Lee, Y.-J., Kang, S.-K. (2004). Design and construction of Korean-Spoken English Corpus. *Proc. of the 8th International Conference on Spoken Language Processing*. pp. 2769—2772.

Strik, H., Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, 29, pp.225—246.

Vertanen, K. (2006). *Baseline WSJ Acoustic Models for HTK and Sphinx: Training Recipes and Recognition Experiments*. University of Cambridge.

Weide, R.L. (2008). *The CMU Pronouncing dictionary*. Online: http://www.speech.cs.cmu.edu/cgi-bin/cmudict/ accessed on 23 Jan 2011.

Weinberger, S.H. (n.d.). *The Speech Accent Archive*. George Mason University. Online: http://accent.gmu.edu/ accessed on 13 Aug 2011.

Witt, S. (1999). Use of Speech Recognition in Computer-Assisted Language Learning. Ph.D dissertation. Cambridge University Engineering Department, UK.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department.