말소리와 음성과학 제3권 제3호 (2011)
pp. 115~121

How Korean Learner's English Proficiency Level Affects English Speech Production Variations

Hong, Hyejin[1)] · Kim, Sunhee[2)] · Chung, Minhwa[3)]

ABSTRACT

This paper examines how L2 speech production varies according to learner's L2 proficiency level. L2 speech production variations are analyzed by quantitative measures at word and phone levels using Korean learners' English corpus. Word-level variations are analyzed using correctness to explain how speech realizations are different from the canonical forms, while accuracy is used for analysis at phone level to reflect phone insertions and deletions together with substitutions. The results show that speech production of learners with different L2 proficiency levels are considerably different in terms of performance and individual realizations at word and phone levels. These results confirm that speech production of non-native speakers varies according to their L2 proficiency levels, even though they share the same L1 background. Furthermore, they will contribute to improve non-native speech recognition performance of ASR-based English language educational system for Korean learners of English.

Keywords: non-native speech recognition, L2 proficiency level, speech production variations

1. Introduction

Current speech recognition systems have attained a level of maturity, and as a result various commercial applications are emerging and becoming more widely adopted. However, non-native speech recognition performance is still considerably lower than native speech recognition performance. This degradation is mainly caused by variations proper to non-native speech [1].

Several studies [2][3][4] have tried to handle variations in non-native speech. All these studies have focused on speakers with a specific L1 background and assume them to be a homogeneous group. However, it is noteworthy that speech production of non-native speakers varies according to their L2 proficiency level, even when they share the same L1 background. One plausible explanation for such variations in speech production can be found in the interlanguage hypothesis [5]: each learner is in his own process of progress toward L2 system, and the process is a dynamic continuum. Thus, it is required to analyze how L2 speech production variations are affected by L2 proficiency levels and to define prototypes which represent the corresponding levels through empirical and quantitative means as well.

In this paper, as a preliminary study to develop an ASR-based English language educational system for Korean learners of English, we examine the relationship between speech production variations and L2 proficiency levels through empirical and quantitative means. A systematic and rigorous analysis of speech production variations at word and phone levels according to L2 proficiency levels are provided by using a large size of speech data.

The remaining part of the paper is organized as follows. Section 2 describes our method for analyzing speech production variations according to L2 proficiency level. Section 3 describes results and discussion about speech production variations obtained

1) Seoul National University, souble1@snu.ac.kr
2) Seoul National University, sunhkim@snu.ac.kr
3) Seoul National University, mchung@snu.ac.kr, corresponding author

Received: August 2, 2011
Revised: August 14, 2011
Accepted: September 17, 2011

by our analysis on a Korean learners' English speech corpus, and conclusions follow in Section 4.

2. Method for analysis of L2 speech production variations

2.1 Speech corpus

To investigate how speech production of L2 learners is affected by their different proficiency levels, ETRI Korean learners' English corpus[4] is used. The corpus consists of 19,883 sentences uttered by 100 native Korean adult learners (48 males and 52 females), aged 20 to 46 years (average of 26.22 years).

Information about learner's English proficiency level is provided with the corpus. Overall English speech proficiency level was evaluated at speaker level on a scale ranging from 1 (poor) to 5 (excellent) by native experts. The learners' proficiency level was determined based on the evaluators' intuition. They did not use any specific quantitative measures to assess learners' speech. Since English proficiency level was not evaluated at utterance level, all utterances spoken by one speaker are labeled as the same proficiency level which corresponds to the speaker's level. For example, all utterances spoken by one speaker of level 5 are of level 5, even though there exist some possibilities that parts of the utterances could be evaluated below level 5. The number of speakers of each speakers' proficiency is given in Table 1.

Table 1. Distribution of L2 proficiency levels

Proficiency level		Number of speakers
1	(poor)	1
2		3
3	↓	30
4		30
5	(excellent)	36

As shown in Table 1, the ETRI corpus shows asymmetry of L2 proficiency level distributions. Only 4 speakers are rated as level 1 or 2, whereas 66 speakers are rated as level 4 or 5. For a comparative analysis, we define two subsets which represent relatively high and low proficiency levels. For a high proficiency level, a subset of the corpus consisting of 2,000 sentences uttered by 20 speakers (10 males, 10 females) of level 5 is selected. The set of a low proficiency level is obtained from level 3, consisting of 2,000 sentences uttered also by 20 speakers (10 males, 10 females). Table 2 shows the statistics of the corpus used for our analysis.

Table 2. Statistics of the corpus used for analysis

	Low proficiency	High proficiency
Number of sentences	2,000	2,000
Number of words	12,009	11,849
Number of vocabulary	2,136	2,062
Number of phones	39,296	38,599

Note that two levels in this paper are not absolute representative of the high and low proficiency level. Speech production variations depend on the data and the learners' proficiency levels defined in them.

2.2 Transcriptions

A total of 77,895 phone tokens are transcribed by 7 transcribers with phonetic knowledge (5 graduate students and 2 undergraduate students in the linguistics department). Standard American English is taken as the norm in the transcription process, since it is the most widely taught in Korean institutions. Transcribers are asked to mark variations which are different from the given canonical pronunciations at phone level.

The initial transcriptions are narrower than phonemic, which allow the transcribers to mark variations by using diacritics. Since our ultimate goal is to improve non-native speech recognition performance, over-expansion of phonetic units needs to be constrained for model construction. Accordingly, more specific phonetic units are collapsed into phoneme-based units, the CMU 39 phoneme set [6] together with a Korean epenthetic vowel [ɯ], which does not exist in the English phoneme inventory [7]. The resulting inter-transcriber agreement is 86.90% which is calculated on 9,327 phones from 498 sentences [8].

2.3 Data analysis

2.3.1 Word-level speech production variations

To investigate how L2 proficiency levels affect speech production, word-level variations are calculated for explaining to what extent speech realizations are different from the canonical forms.

Word-level performance is measured by correctness (%), which indicates the percentage of words that are matched with the canonical pronunciation forms. Since our focus is on pronunciation variations, not on language use or reading competence, inserted or deleted words are not included. For this reason, deletion and insertion are not taken into account in the following formula of correctness[5].

4) This corpus is a collection of English speech uttered by Korean adults. A part of the corpus was used for this study, and the corpus is not publicly available.

$$Correctness(\%) = \frac{N-S}{N} \times 100 \quad (1)$$

(N: total number, S: substitution)

Phonetic realizations of each word are compared across different proficiency levels.[6] At first, all variants are compared and classified into two major categories: common variants and level-dependent variants, as shown in Figure 1. Phonetic variants for the same word which occur in speech of both levels are common variants, whereas level-dependent variants occur in the speech of one level only. For example, one pronunciation variant /d eh m/ for the word 'THEM' occur in speech of both levels, while another variant /dh ah m/ occur only in the speech of the high proficiency level learners. It is a common variant in the former case, while the latter is a level-dependent variant. All variants with the given categories are compared with the canonical forms.

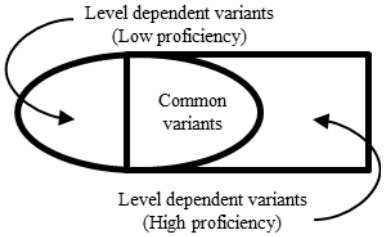

Figure 1. Two major categories of variants: common variants and level-dependent variants. Oval-shaped area indicates phonetic variants space for the low proficiency level, whereas rectangle-shaped area is for the high proficiency level. The overlapped area is for common variants.

2.3.2 Phone-level speech production variations

L2 proficiency level-dependent variations in speech production are analyzed at phone level as well.

L2 learners show learning errors caused by negative transfer from L1, called interlingual errors [9]. Non-native speakers tend to substitute L2 phonemes with L1 phonemes. In addition, they insert or delete phonemes, especially when L2 syllable structure is not permitted in L1 [7]. This leads to use a different measure from the one for word-level variation, which will be the accuracy (%) including insertions and deletions together with substitutions.

$$Accuracy(\%) = \frac{N-D-S-I}{N} \times 100 \quad (2)$$

(N: total number, D: deletion, S: substitution, I: insertion)

Detailed phone level analysis is performed based on confusion matrix, which shows information about which phone is confusing and how it is realized. The confusion matrices are generated by calculating how many times a phone instance is realized as the target phone.

The variation of individual phones in terms of accuracy is compared according to the proficiency level to scrutinize the difference between the speech production variations of two proficiency levels. Certain phones with the largest difference between two proficiency levels are decided for more detailed analysis.

3. Results and discussion

3.1 Word-level speech production variations

The results of analysis on word-level variations are provided in Table 3.

Table 3. Speech production variations at word level

		Low proficiency	High proficiency
Overall performance correctness (%)		67.89	73.78
Common variants	Type	1,188	
	Matched (with canonical)	785	
	Unmatched	403	
	Token	9,421	9,743
	Matched	7,469	7,679
	Unmatched	1,952	2,064[7]
Level-dependent variants	Type	1,041	742
	Matched	71	182
	Unmatched	970	560
	Token	1,361	953
	Matched	113	282
	Unmatched	1,248	671
Variants per word		1.96	1.70

For overall performance, speakers with the high proficiency level outperform speakers of the low proficiency level by 5.89%.

5) As a result, it is different from the general correctness measure used for measuring speech recognition performance.

6) In our speech corpus, sentences uttered by each speaker are different. Thus, a set of words that commonly occur in both proficiency levels' speech are chosen, and then word-level variation analysis is performed on this set.

7) More unmatched common variants for the high proficiency level means that unmatched variants are the common variants rather than the level-dependent variants. 75.47% of the unmatched variants of the high proficiency level are the common variants, while 61.00% of the unmatched variants are common in the case of the low proficiency level.

This performance difference is statistically significant[8] (P<0.0001). These results mean that the speech production of the high proficiency level is more similarly realized to the canonical forms than that of the low proficiency level.

Detailed. analysis on word-level variations is presented based on variants' types and tokens. It is shown that more word-level pronunciation variants occur in speech of the learners with the low proficiency level. The number of variants per word is 1.96 for the low proficiency level in average, whereas each word entry has 1.70 pronunciation variants in the case of the high proficiency level. However, occurrence of more pronunciation variants does not mean that all variants found in speech of the low proficiency level can cover that of the high proficiency level. While some of the pronunciation variants of the low proficiency level overlap with that of the high proficiency level, both have their own proficiency level-dependent variations. The low proficiency level-dependent variants are 1,041 types, which is 46.70% of the variants. In the case of the high proficiency level, 742 types occur as the level-dependent variants, which is 38.45% of all variants.

We calculate how many level-dependent variants are realized as the canonical forms. Among the level-dependent variants, 24.53% of the high proficiency level-dependent variants are found to be the same as the canonical forms, whereas only 6.82% of the low proficiency level-dependent variants are matched with the canonical forms. These results indicate that the learners with the high proficiency level are in the process of progress toward more L2-like system than the low proficiency level learners in terms of matching with the canonical forms. One crucial factor which leads to poorer word-level performance of the low proficiency level can be found in these results.

There exists another factor to decrease the performance of the low proficiency level. It is worthy to examine the frequency rates of tokens in the case of the common variants, especially when the common variants are the same as the canonical forms. An example of pronunciation variants and their frequency rates in different levels are provided in Table 4.

Table 4. An example of pronunciation variants for the word 'VERY' and their frequency rates (%) in low proficiency and high proficiency

		Low proficiency	High proficiency
VERY (canonical)	V EH R IY /v e r i/	46.67	88.24
VERY (1)	B EH R IY /b e r i/	40.00	11.76
VERY (2)	V EH IY /v e i/	3.33	-
VERY (3)	W EH R IY /w e r i/	3.33	-
VERY (4)	B EH L IY /b e l i/	3.33	-
VERY (5)[9]	B EH R IH /b e r ɪ/	3.33	-

As presented in Table 4, both proficiency levels have different frequency rates for the canonical forms. More tokens are realized as the canonical forms in speech of the high proficiency level than that of the low proficiency level. For example, 88.24% of the pronunciation variants for the word 'VERY' are realized as the canonical form in the case of the high proficiency level. However, only 46.67% of tokens for the word are decided to be the same as the canonical forms in the low proficiency level. These imply that a smaller portion of the pronunciation variants is judged as the canonical forms in the case of the low proficiency level, which leads to performance degradation.

3.2 Phone-level speech production variations

To analyze L2 proficiency level-dependent variations, the performance of speech production is measured in terms of accuracy at phone level as well. The results of the overall performance are presented in Table 5.

Table 5. Overall performance at phone level

Accuracy (%)	Low proficiency	High proficiency
Phone level	86.93	90.01

As expected, the same results are obtained at phone level as well; the learners with the high proficiency level show higher performance by 3.08%, which is statistically significant difference

8) All statistical significance hereafter presented in this paper is based on a Chi-square test for the comparison of two proportions at a level < 0.05.

9) One reviewer pointed out that the distinction between VERY (1) and VERY (5) is not meaningful, since the final vowel is actually lax vowel and the transcribers did not seem to be able to distinguish two vowels (lax/tense). However, two variants are separately presented according to the original transcription.

(P<0.0001).

All individual phone performance is inspected as shown in Figure 2, and detailed phone level analysis based on confusion matrix analysis is performed.

Criteria to decide phones to be analyzed in detail are set based on both phone performance and frequency. First, all phones whose correctness difference between two proficiency levels is more than 2.34% (overall phone correctness difference) are selected. Relative difference is considered as well; phones with relative difference of less than 50% are excluded. Finally, phones which occur more than 200 times are selected. Considering phone frequency together with performance seems to be reasonable, since insufficient occurrences are not representative and reliable for analysis. Moreover, if one phone does not occur frequently, its impact on performance may not be considerable. For example, 'ZH' which shows 14.29% performance difference is not included, since its frequency is 15 and 14 in the low and high proficiency speech respectively. As a result, six phones are chosen as the final list as shown in Table 6.

Table 6. Comparisons of individual phone performance in accuracy (%): 6 phones with the largest performance difference are shown.

Phones	Low proficiency	High proficiency	Performance difference
V /v/	70.69	93.16	22.47 (P<0.0001)
TH /Θ/	68.80	86.52	17.72 (P<0.0001)
Z /z/	69.17	85.36	16.19 (P<0.0001)
ER /ɝ/	68.47	84.38	15.91 (P<0.0001)
F /f/	82.53	95.32	12.79 (P<0.0001)
R /r/	76.88	84.69	7.81 (P<0.0001)

These phones have one common point: they do not exist in the Korean phoneme inventory. The large performance difference for these phones means that the low proficiency level learners have more difficulties in producing phones which are not in L1 phoneme inventory than the high proficiency level learners do. In Table 7, six phones which show the largest performance difference and their realizations according to L2 proficiency level are presented.

Table 7. Target phones (6 phones with the largest performance difference) and their realizations with corresponding substitution rates (%)[10]. All phones presented in the table show statistically significant performance difference.

Target phones	Realized phones	Low proficiency	High proficiency	Performance difference
V /v/	B /b/	17.94	4.27	13.67 (P<0.0001)
TH /Θ/	S /s/	21.43	3.48	17.95 (P<0.0001)
Z /z/	S /s/	27.38	12.76	14.62 (P<0.0001)
Z /z/	del	2.40	0.84	1.56 (P=0.0130)
ER /ɝ/	AH /ʌ/	16.11	6.00	10.11 (P<0.0001)
F /f/	P /p/	14.64	3.78	10.86 (P<0.0001)
R /r/	L /l/	8.48	2.86	5.62 (P<0.0001)
R /r/	del	5.82	3.60	2.22 (P=0.0025)

In the case of the low proficiency level learners, 17.94% of 'V'(/v/) are realized as 'B'(/b/), whereas only 4.27% of 'V' are substituted with 'B' in the high proficiency level learners' speech.

The low proficiency level learners substitute 21.43% of English voiceless interdental fricative 'TH'(/Θ/) by voiceless alveolar fricative 'S'(/s/). However, only 3.48% of 'TH' are replaced by 'S' in the case of the high proficiency level learners. Instead, they tend to substitute 'D'(/d/) for 'TH'. In the speech of the high proficiency level learners, 7.83% of 'TH' are realized as 'D'. Substitution rate of 'TH' with 'D' is 4.14% in the case of the low proficiency level learners. This means that strategies that are utilized for overcoming problematic phones are differently compiled according to the learners' proficiency level.

Voiced alveolar fricative 'Z'(/z/) seems to be more problematic for the low proficiency level learners than the learners with high proficiency level. 27.38% of 'Z' are realized as voiceless alveolar fricative 'S'(/s/) in the speech of the low proficiency level learners. For the high proficiency level learners, considerably lower rates, 12.76% of 'Z' are substituted with 'S'. Deletions of 'Z' are more frequently found in the speech of the low

10) One of the reviewers provided very important comments; Substitution of 'Z' with 'S' has to be considered as a phonetic variation if the condition is word-final position. In the case of 'ER' and 'AH', it is appropriate to consider difference of two vowels as a phonetic variation. More detailed analyses should be followed to investigate which phonetic variations are acceptable or occur in native speech as well, accordingly plausible or proper phonetic realizations.

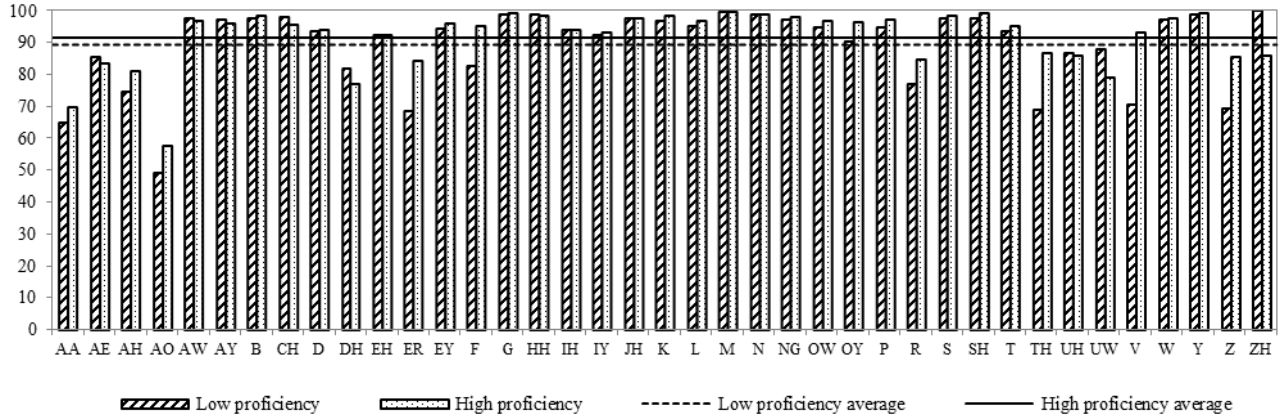

Figure 2. Phone-level performance of speech production according to two proficiency levels. Bar graph shows the performance of individual phones assessed using phone correctness (%). Line graph presents the average phone-level performance of all phones using phone accuracy (%) which considers insertions and deletions together with substitutions.

proficiency level learners as well. 2.40% of 'Z' deletions are found in the low proficiency level learners' speech, while only 0.84% are deleted in the speech of the high proficiency level learners.

English rhoticized vowel 'ER'(/ɝ/) is found to be more difficult for the low proficiency level learners. They tend to replace it by 'AH'(/ʌ/), and the substitution rate is 16.11% which is higher by 10.11% than that of the high proficiency level learners.

Another finding is that the low proficiency level learners have difficulties in producing voiced labiodental fricative 'F'(/f/). 14.64% of 'F' are produced as bilabial stop 'P'(/p/). This substitution is found in the speech of the high proficiency level learners as well, however, the rate remains relatively low, 3.78%.

Substitutions of 'R'(/r/) by 'L'(/l/) are found in both the high proficiency level and the low proficiency level learners. However, the rate is much higher in the low proficiency level learners (8.48% vs. 2.86). 5.82% of 'R' are deleted in the speech of the low proficiency level learners and 3.60% in that of the high proficiency level learners.

As discussed, the low proficiency level learners tend to realize target phones as other phones which differ in terms of manner or place of articulation with a higher proportion of substitution rates.

This implies that the degree of L1 interference in speech production does not affect the learner equally, but rather depends on their proficiency level. These results reinforce that the learners with the high proficiency level have more L2-like system than the learners with the low proficiency level.

3.3 L2 proficiency level and ASR performance

Given that the recognition performance is degraded in non-native speech, speech recognition experiments are performed to show that there is a correlation between L2 proficiency level and ASR performance.

The recognizer used in the experiment is implemented using HTK v.3.4 [10]. An acoustic model is constructed following the process provided in [11]. The number of Gaussians is increased up to 16 for phones (32 for silence). The CMU pronunciation dictionary is used for the lexicon. The statistics of the corpora used in our experiments is shown in Table 8.

Table 8. Statistics of the corpora used in experiments

Corpus			Sentence
Training		WSJ0/ WSJ1	101,635
Test	Native	Nov92	330
	Non-native	Low proficiency	2,000
		High proficiency	2,000

Table 9 shows the experimental results of speech recognition.

Table 9. ASR performance

Test	Accuracy (%)
Low proficiency	42.96
High proficiency	72.08
Native	94.88

The performance rate of both high performacne and low performance of non-native speech is lower than that of native speech (94.88%). And, the performance of the high proficiency level is 29.12% higher than that of the low proficiency level. Despite this considerable performance gap, note that all learners have the same L1 language, Korean. From this, we expect that

some potential improvement in ASR performance may be obtained if speech production variations are appropriately modeled according to L2 proficiency levels.

4. Conclusions

This paper examines how speech production variations are affected by L2 proficiency level, realizing that speech production variations of non-native speakers degrade ASR performance. It focuses on a systematic and detailed analysis of speech production variations at word level and phone level using the Korean learners' English corpus. The results of the analysis show that speech production of learners with different L2 proficiency levels are considerably different in terms of overall performance and the individual realizations at word and phone levels. These results confirm that speech production varies even in a group of learners with the same L1 background. The details of the analysis results can be used to improve non-native speech recognition performance of ASR-based English language educational system for Korean learners of English.

In our future research, more detailed analyses on relationships between learners' proficiency levels and speech production variations will be performed using different speech data. From this comparative research, criteria to generate and select phonetic variants according to L2 proficiency levels will be achieved.

Acknowledgements

This work was supported by the Industrial Strategic Technology Development Program, 10035252, "Development of dialog-based spontaneous speech interface technology on mobile platform," funded by the Ministry of Knowledge Economy (MKE, Korea).

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*, Vol. 49, No. 10-11, pp. 763-786.

Tomokiyo, L. M. (2000). "Lexical and acoustic modeling of non-native speech in LVCSR", *Proceedings of ICSLP 2000*, pp. 346-349.

Goronzy, S., Rapp, S. & Kompe, R. (2004). "Generating non-native pronunciation variants for lexicon adaptation", *Speech Communication*, Vol. 42, No. 1, pp. 109-123.

Bouselmi, G., Fohr, D., Illina, I. & Haton, J. P. (2006). "Fully automated non-native speech recognition using confusion-based acoustic model integration and graphemic constraints", *Proceedings of ICSLP 2006*, pp. 345-348.

Selinker, L. (1972). "Interlanguage", *International Review of Applied Linguistics*, Vol. 10, pp. 209-231.

Carnegie Mellon University, *The CMU Pronouncing Dictionary,* http://www.speech.cs.cmu.edu/cgi-bin/cmudict, accessed on 23 Jan 2011.

Hong, H., Kim, J. & Chung, M. (2010). "Effects of Korean learners' consonant cluster reduction strategies on English speech recognition performance", *Proceedings of INTERSPEECH 2010*, pp. 1858-1861.

Ryu, H., Lee, K., Kim, S. & Chung, M. (2011). "Improving transcription agreement of non-native English speech corpus transcribed by non-natives", *Proceedings of SLaTE 2011*.

Richards, J. C. (1971). "A non-contrastive approach to error analysis", *English Language Teaching*, Vol. 25, pp. 204-219.

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.

Vertanen, K. (2006). "Baseline WSJ acoustic models for HTK and Sphinx: training recipes and recognition experiments", Technical report, University of Cambridge.

• **Hong, Hyejin**
Department of Linguistics
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea
Tel: 82-2-880-9039
Email: souble1@snu.ac.kr

• **Kim, Sunhee**
Center for Humanities and Information
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea
Tel: 82-2-880-7735
Email: sunhkim@snu.ac.kr

• **Chung, Minhwa**, corresponding author
Department of Linguistics
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea
Tel: 82-2-880-9195
Email: mchung@snu.ac.kr