Modeling Cross-morpheme Pronunciation Variations for Korean Large Vocabulary Continuous Speech Recognition

Kyong-Nim Lee and Minhwa Chung

Department of Computer Science, Sogang University
Sinsu-Dong, Mapo-Ku, Seoul 121-742, Korea
{knlee, mchung}@sogang.ac.kr

Abstract

In this paper, we describe a cross-morpheme pronunciation variation model which is especially useful for constructing morpheme-based pronunciation lexicon for Korean LVCSR. There are a lot of pronunciation variations occurring at morpheme boundaries in continuous speech. Since phonemic context together with morphological category and morpheme boundary information affect Korean pronunciation variations, we have distinguished pronunciation variation rules according to the locations such as within a morpheme, across a morpheme boundary in a compound noun, across a morpheme boundary in an *eojeol*, and across an *eojeol* boundary. In 33K-morpheme Korean CSR experiment, an absolute improvement of 1.16% in WER from the baseline performance of 23.17% WER is achieved by modeling cross-morpheme pronunciation variations with a context-dependent multiple pronunciation lexicon.

1. Introduction

It is well known that reflecting pronunciation variations into the lexicon improves the CSR's performance. Usually the baseline lexicon contains a single pronunciation per each entry. However, due to coarticulation a lot of the pronunciation variations occur across boundaries of lexical entries [1][2].

A morpheme is defined as the minimal grammatical unit or the minimal meaningful unit of language [4]. The spacing unit in Korean orthography is called *eojeol*, which results from combining substantial and formal morphemes. Since Korean is an agglutinative language, *eojeol* has different properties from the spacing unit, *word* in English. Hundreds of *eojeols* can be easily generated from a given root word by combining substantial and formal morphemes. Since it is impractical to use all the combinations of morphemes as lexical entries of pronunciation lexicon, most Korean LVCSR systems choose morphemes as basic analysis units for lexical and language models.

One problem in using morphemes as lexical units for CSR systems is that many morphemes have two or more different realizations, called allomorphs, determined by the final sound in the stem to which they are added. Furthermore, lexical variations for nouns are appeared to be particular due to compounding or phrase words in Korean. In English, *multi-words* [5] are used to provide a solution combining the ease of modeling at the lexical level with the need to model crossword variations. For similar reasons, sound-based *pseudo-morphemes* and *multi-morphemes* obtained by concatenating morphemes are widely used for Korean CSR[8].

In a multiple-morpheme-based pronunciation lexicon, morpheme boundaries are included in a multi-morpheme entry resulting from compound nouns or concatenations of formal morphemes such as particle, suffix, and predicate ending. In most cases, typically in "tensification", "liaison", "/*d*/-palatalization", and "/*n*/-insertion", Korean pronunciation variations across morpheme boundaries are different from the variations occurring within a morpheme, although they have the same phonemic context. Cross-morpheme variations need to be modeled differently from within-morpheme variations.

This paper describes a cross-morpheme pronunciation variation model which is especially useful for constructing a morpheme-based pronunciation lexicon for Korean LVCSR. Phonemic context together with morphological category and morpheme boundary information affect Korean pronunciation variations. To address this problem, we have distinguished pronunciation variations based on the locations such as (1) within a morpheme, (2) across a morpheme boundary in a compound noun, (3) across a morpheme boundary in an *eojeol*, and (4) across an *eojeol* boundary. Our pronunciation variation model produces a range of possible phonetic transcriptions according to phonemic contexts and morphological properties of an input orthographic transcription. By analyzing phonological variations frequently found in spoken Korean, we have derived 816 pronunciation rules based on phonemic contexts and morphological information. We employ the context-dependent multiple pronunciation lexicon, which reflects the pronunciation variations across morpheme boundaries. Our experiments show that modeling cross-morpheme pronunciation variation in the pronunciation lexicon improves the CSR's performance.

This paper is organized as follows. Section 2 introduces the characteristics of Korean pronunciation variations. Section 3 describes our model for handling cross-morpheme pronunciation variations. In section 4 and 5, we describe the task material and baseline system, followed by experimental results and conclusions.

2. Korean Pronunciation Variations

Based on literature survey [6], we have identified 20 major phonemic rules [3], which explain the phonemic variations frequently found in spoken Korean. We have also derived 816 phonemic contexts, which can trigger the application of the corresponding phonemic rules. In most cases, a phonemic context is defined as an ordered set of two adjacent consonants: the final consonant of a syllable and the first consonant of the next syllable.

The following discussion shows one of the representative phonemic variations in Korean.

Tensification: In Korean, stops in the final consonant position generally become tense or geminated before another obstruent. A lenis obstruent onset following any obstruent

coda at the morpheme boundary in a compound word is pronounced as a fortis obstruent. The application of tensification at a morpheme boundary glues together the components in a sequence. Tensification is categorized as follows [6]:

- Post-obstruent Tensification: For example, "*beob-dae*" (meaning "law school") → /*beob-**tt**ae*/ (the corresponding phonemic transcription)
- Predicate Tensification: For example, "*sin-go*" (meaning "wearing") → /*sin-**kk**o*/ "*gam-gi*" (meaning "winding") → /*gam-**kk**i*/
- Modifier Tensification: When the determiner (or modifier derived from a verb/adjective) "*l*" of an *eojeol* is followed by another *eojeol* with a non-unit bound noun, the initial consonant of the second *eojoel* is tensified. For example, "*gal-ji*" (meaning "whether to go") → /*gal-**jj**i*/

Regarding the examples for predicate tensification, if the same phonemic configurations occur not at a morpheme boundary but within a morpheme, tensification does not happen. In this case, the resulting phonemic transcriptions are /*sin-go*/ for the input "*sin-go*" (meaning "report") and /*gam-gi*/ for "*gam-gi*"(meaning "cold"). In other words, the phonemic configurations and morphological properties of an orthographic transcription affect pronunciation variations. For example, an *eojeol* "g*am-gi*" may be a noun or a combination of an *eogan*[1] "g*am*" and an *eomi*[2] "g*i*". When "g*am-gi*" is a noun, its phonemic and phonetic transcriptions are /*gam-gi*/ (meaning "cold") and /K AA M G IY/, respectively. When it is a combination of an *eogan* and an *eomi*, its phonemic and phonetic transcriptions are /*gam-kki*/ (meaning "winding") and /K AA M KK IY/, respectively. That is, the first consonant "g" of the second syllable "g*i*" is pronounced as either /G/ or /KK/ depending on the morphological properties. The phonemic configuration also affects pronunciation. The first consonant "g" of the first syllable "g*am*" would be pronounced as /G/ if the preceding phoneme is a voiced sound. Therefore, we have 4 possible pronunciations for the word "g*am-gi*".

Furthermore, some pronunciation variations such as "vowel shortening of predicative conjugated form", "/*n*/-insertion", "final consonant /*h*/-deletion", and "tensification", are realized differently depending on the morpheme categories such as noun, verb, and particle.

3. Modeling Cross-morpheme Pronunciation Variations

3.1. Lexical Level Pronunciation Variations

To incorporate a pronunciation variation model at the lexical level, we use a multiple pronunciation lexicon, which includes alternative pronunciations for each entry. Given an acoustic evidence A, the goal of speech recognition is to find the word sequence W that maximizes the probability P(W|A). According to the Bayes' rule, we have

$$\hat{\mathrm{W}} = \arg\max_{\mathrm{W}} P(\mathrm{W})P(\mathrm{A}\,|\,\mathrm{W}) \qquad (1)$$

P(W) is given by language model and the acoustic likelihood P(A|W) is computed from acoustic model. If pronunciation variations are taken into account, Equation (1) is modified into Equation (2), where $\mathrm{L}_{\mathrm{W},k}$ is the *k*-th pronunciation variant for the word W. The modified equation essentially searches for particular pronunciation variants that maximize the probability. $\mathrm{P}(\mathrm{A}|\mathrm{L}_{\mathrm{W},k})$ is the acoustic likelihood of pronunciation $\mathrm{L}_{\mathrm{W},k}$.. $\mathrm{P}(\mathrm{L}_{\mathrm{W},k}|\mathrm{W})$ gives the probability that W is pronounced as $\mathrm{L}_{\mathrm{W},k}$.

$$\hat{\mathrm{W}} = \arg\max_{\mathrm{W},k} P(\mathrm{W})P(\mathrm{A}\,|\,\mathrm{L}_{\mathrm{W},k})P(\mathrm{L}_{\mathrm{W},k}\,|\,\mathrm{W}) \quad (2)$$

An important distinction that is often drawn in pronunciation modeling is between within-morpheme and cross-morpheme pronunciation variations. Within-morpheme pronunciation variations can be easily generated based on phonemic configurations. They are usually encoded as the baseform pronunciation in the lexicon. On the contrary, cross-morpheme pronunciation variations depend upon not only phonemic configurations but also neighboring morphemes' categories. Table 1 shows how the baseform pronunciation /K JO JU KQ/ of noun "*gyo-yuk*" (meaning "education") generates different pronunciations depending upon the neighboring morphemes' categories and phonemic configurations.

Table 1. Various phonetic transcriptions of "*gyo-yuk*"

Phonetic transcription of "*gyo-yuk*"	Examples of morpheme concatenations	$\mathrm{P}(\mathrm{L}_{\mathrm{W},k}\|\mathrm{W})$
K JO JU G	*gyo-yuk* + *i*	0.46222
K JO JU KQ	*gyo-yuk* + *gwa*	0.40889
G JO JU KQ	*sa-hoe* + *gyo-yuk*	0.04444
G JO JU G	*cham* + *gyo-yuk* + *i*	0.03333
K JO JU KH	*gyo-yuk* + *hae*	0.02889
K JO JU NX	*gyo-yuk* + *man*	0.00667
KK JO JU G	*dae-hak* + *gyo-yuk* + *i*	0.00667
KK JO JU KQ	*dae-hak* + *gyo-yuk*	0.00667
G JO JU NX	*ui-mu* + *gyo-yuk* + *man*	0.00222

3.2. Cross-morpheme Pronunciation Variations

For grapheme-to-phoneme conversion, cross-morpheme letter sequences are a major source of ambiguity. In order to encode cross-morpheme pronunciation variations into pronunciation lexicon, our model generates them by applying phonemic change rules based upon morpheme boundary information.

There are some important observations in describing cross-morpheme pronunciation variations in Korean: (1) Cross-morpheme pronunciation variations are clearly different from within-morpheme pronunciation variations. The same phonemic configurations often correspond to different pronunciations depending on their location. (2) Cross-morpheme pronunciation variations in compound nouns are particularly different from those in other morphological categories. (3) Cross-morpheme pronunciation variations at an *eojeol* boundary within an *eonjeol* are limited to the following phonemic changes such as "tensification", "/*n*/-insertion", "liaison", "aspirationalization", "nasalization of obstruent", and "conversion into bilabial or velar sound". Other pronunciation variations do not happen at an *eojeol*

[1] *Eogan* is a root of a verb or adjective in Korean.
[2] *Eomi* is a verb-ending or adjective-ending in Korean.

boundary within an *eonjeol*. *Eonjeol* is a phonological phrase or a unit of pause in spoken Korean. More than one *eojeol* are often uttered without pause to form an *eonjeol* [6]. Therefore, applications of cross-morpheme pronunciation rules depend on whether the location is at an *eojeol* boundary or within an *eojeol*.

Based on these observations, we have classified cross-morpheme pronunciation variations depending on the location of the cross-morpheme as follows:

- Between morphemes in an *eojeol*
  - *noun + subject particle:*
    gyo-yuk+i (meaning "education is") → *gyo-y**u**+**g**i*
    som+i (meaning "cotton is") → *s**o**+**m**i*
  - *eogan + eomi:*
    sin+da (meaning "wear") → *sin+**tt**a*
- Between morphemes in a compound noun
  - *noun + noun* in a *compound noun:*
    som+i-bul (meaning "cotton sheet") → *som+**ni**-bul*
- Between morphemes at *eojeol* boundary in an *enjeol:*
  - *noun # noun:*
    dae-hak # gyo-yuk (meaning "university education") → *dae-hak # **kk**yo-yuk*
  - *adnominal verb-ending # non-unit bound noun:*
    l # su (meaning "able to") → *l # **ss**u*

Here, '-' stands for a syllable boundary, '+' for a morpheme boundary, and '#' for an *eojeol* boundary, respectively.

3.3. Modeling Cross-morpheme Pronunciation Variations

This idea is applied to the grapheme-to-phoneme conversion in our pronunciation variation generator [3]. It produces a phonetic transcription by using a set of finite state automata, which transforms an underlying lexical representation to its surface representation. Here, the underlying representation corresponds to an orthographic transcription and the surface representation corresponds to a phonetic transcription.

Our pronunciation generation algorithm goes through the following steps: (1) morphological analysis of an input orthographic transcription, (2) applications of 15 obligatory phonemic rules according to morphological categories to generate a canonical phonemic transcription, (3) applications of 5 optional phonemic rules to generate a non-standard phonemic transcription that happens frequently in spoken Korean, and (4) applications of 3 major allophonic rules to generate a phonetic transcription.

We have modeled 20 major phonemic changes using 816 subrules, as shown in Table 2. Since phonemic variations depend on both morphological category and phonemic context, we have augmented phonemic rule automata with the scope of the rule application such as 'E/M/C/I/O/m' flags. 'E' means that the rule can be applied to *eojeol* boundary, 'M' to morpheme boundary, 'C' to morpheme boundary of a compound noun, and 'I' to any location within a morpheme, respectively. 'O' stands for an optional rule and 'm' for a multiply applicable rule, respectively.

In Table 2, L3 stands for the last consonant of a syllable, R1 for the first consonant of the following syllable, rule #4 for "liason", rule #9 for "tensification", and rule #14 for "conversion into bilabial or velar sound", respectively. Rule #14 is an optional rule and can be applied to any location. By applying these rules, we can get the right phonemic transcription /*sin-ba+reul sin-kko*/ from an input text "*sin-bal+eul # sin-go*" by applying rules 4.8 and then 14.1 to /*sim-bal+eul*", and rules 9.133 and then 14.9 to /*sing-kko*/, respectively.

Table 2. Examples of phonemic change rules.

Phonemic Context			Change Code		Rule No.	Scope
L3	R1		L3	R1		E/M/C/I/O/m
l	Ø	→	Ø	*r*	4.8	1 1 0 0 0 0
n	*g*	→	*n*	*kk*	9.133	0 1 1 0 0 1
n	*b*	→	*m*	*b*	14.1	-
n	*kk*	→	*ng*	*kk*	14.9	-

4. Experimental Results

4.1. Task Material and Baseline

Acoustic models for speech recognition have been designed and constructed for common use on sentences that cover various phonological environments and balanced triphones. For this purpose, 60K phonetically balanced sentences with 44K Korean morphemes are collected from newspapers and textbooks. The corresponding speech data lasted 100 hours of recoding time and were produced by 600 speakers.

Table 3. Statistics of the Samsung PBS DB[1].

Count	# Total	# Unique	Average
Sentence	60,000	60,000	9.2 *eojeols*
Tagged *eojeol*	551,820	170,419	2.1 morphemes
Tagged morpheme	1,160,597	44,303	1.9 syllables
Syllable	2,230,845	167,949	-
Morpheme boundary	608,777	-	-

The speech signal was sampled at 16KHz and segmented into 25ms frames with each frame advancing every 10ms. Each frame was parameterized by MFCC-based 39-D feature vector that consists of 12 MFCC parameters and their differential coefficients of 1st and 2nd order, together with power, and its corresponding time derivatives. Each of the sets is modeled by tied-state triphone CHMM. Each HMM is a left-to-right model of five states, each with six mixtures and has been trained and tested using the Baum-Welch and Viterbi algorithms [7]. We have used 43K utterances for training, 58K sentences for language modeling, and 686 utterances without OOV for testing. In our experiment, the lexicon contains less than 40K unique entries. Back-off morpheme bigram is used for language model, which perplexity is 120.87 and network size is 370K (6.92 bits of entropy and 87.88% of bigram hit ratio).

4.2. Performance of Pronunciation Variation Models

The Samsung PBS database is orthographically annotated at the word level. Since there is no standard large vocabulary Korean pronunciation lexicon available, we have used a speech recognizer to make a large vocabulary pronunciation lexicon. As a reference, handcrafted auditory transcriptions for 43K sentences are used to make an initial acoustic model.

[1] HCI Lab of Samsung Advanced Institute of Technology has funded our construction of the PBS DB.

For more accurate training, acoustic models are refined by forced Viterbi alignment using pronunciation lexicon containing alternative pronunciations. Then the resulting phonetic sequence is used to constrain an optimal alignment between existing acoustic models and the speech data.

We used two criteria for performance evaluation. One is PER (phone error rate), which is the result of the comparison between auditory transcriptions and phonetic transcription after forced alignment. The other is WER (word error rate), which is actually a morpheme error rate. The baseline pronunciation lexicon contains only those within-morpheme pronunciation variations, which are derived using phonemic contexts only. Therefore, the baseline lexicon contains context-dependent alternative pronunciations, but does not reflect pronunciation variations at morpheme boundaries. Data 2 in Table 4 shows that classifying phonemic contexts further using part-of-speech information does not significantly improve the performance.

Table 4. Performance of pronunciation variation models: WM stands for within-morpheme variations, POS for part-of-speech information, and XM for cross-morpheme variations, respectively.

	Pronunciation Lexicon	# of entries	PER (%)	WER (%)
1	Baseline (WM)	33,398	84.91	23.17
2	WM + POS	33,400	84.97	23.16
3	WM + weak XM (the same rules for WM and XM)	39,733	93.42	22.42
4	WM+XM	39,722	93.46	22.04
5	WM + XM + POS	39,725	93.49	**22.01**
6	WM + XM + POS + splitting nouns	37,398	92.76	22.08
7	WM + XM + POS + modifier tensification	39,735	93.50	22.14

Data 3 to 7 in Table 4 show the results of reflecting cross-morpheme pronunciation variations. Data 3 indicate that modeling cross-morpheme variations based on phonemic contexts by using the same rules that are used for within-morpheme variations produces an absolute improvement of 0.74% in WER and 8.51% in PER. This implies the usefulness of a cross-morpheme pronunciation variation model. As shown in Data 4, when we applied different rules to model within-morpheme and cross-morpheme variations, an absolute improvement of 0.38% in WER from Data 3 is achieved. This means distinguishing cross-morpheme variations from within-morpheme variations is of the utmost importance. In this study, we found out that these are actually 47 phonemic contexts having different rules among all rules. Data 5 shows that the best performance in WER is achieved when POS information is used together with a cross-morpheme pronunciation variation model. Data 6 explains that splitting compound nouns produces no improvement in WER except a reduction in lexicon size. Data 7 shows that applying 'modifier tensification' compulsorily at *eojeol* boundary produces the best PER, but leads to a deterioration in WER, because the added variants can be confused with other lexical entries.

In summary, compared to the baseline performance, 1.16% reduction of WER is obtained by a cross-morpheme pronunciation variation model with a phonemic-context-dependent multiple pronunciation lexicon, whereas 8.59% improvement in PER is obtained by a cross-morpheme pronunciation variation model applying 'modifier tensification' compulsorily at *eojeol* boundaries.

5. Conclusions

We have described a cross-morpheme pronunciation variation model which is especially useful for constructing morpheme-based pronunciation lexicon for Korean LVCSR. The best result was obtained when we distinguished cross-morpheme variations from within-morpheme variations and used a phonemic-context-dependent multiple pronunciation lexicon. However, we found out that adding pronunciation variants to the lexicon usually introduces new errors because the confusability within the lexicon increases. For this reason, optimal selection of pronunciation variants is important. An obvious criterion for variants selection is the frequency of occurrence. Besides absolute and relative frequency, many selection criteria such as entropy and likelihood can be used [1]. To further dwell this topic, we intend to deal with variants selection approach to improve the performance of Korean LVCSR.

Acknowledgements

This work has been supported by Ministry of Science and Technology's Brain Neuroinformatics Research Program (Project No. M1-0107-01-0003). We are grateful to HCI Lab of SAIT for providing the read speech DB for our experiments.

References

[1] H. Strik, C. Cucchiarini, "Modeling Pronunciation Variation for ASR: A Survey of literature," *Speech Communication*, 29(2-4): pp. 225-246, 1999.

[2] J. Kessens, M. Wester and H. Strik, "Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation Variation," *Speech Communication*, 29(2-4): pp. 193-207, 1999.

[3] K.-N. Lee, J. Jeon and M. Chung, "Automatic Generation of Pronunciation Variants for Korean Continuous Speech Recognition," *The Journal of the Acoustical Society of Korea*, 20(2): pp. 35-43, 2001.

[4] L. Bauer, *Introducing Linguistic Morphology*, Edinburgh University Press, 1988.

[5] M. Finke and A. Waibel, "Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition," *Proceeding of the 5th European Conference on Speech Communication and Technology*, pp.2379-2382, 1997.

[6] S.-C. Ahn, *An Introduction to Korean Phonology, Hanshin Press*, 1998.

[7] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtcher, P. Woodland, *The HTK Book* (for HTK Version 2.2), Entropic Cambridge Research Laboratory, 1999.

[8] Y.-H. Park and M. Chung, "Automatic Generation of Concatenate Morpheme Based Language Models for Korean LVCSR," *Proceedings of the International Conference on Speech Processing*, pp.633-637, 2001.