



# Utilizing a Noisy-Channel Approach for Korean LVCSR

Sakriani Sakti, Ryosuke Isotani, Hisashi Kawai, Satoshi Nakamura

Spoken Language Communication Research Group, MASTAR Project,  
National Institute of Information and Communications Technology (NICT), Japan

{sakriani.sakti, ryosuke.isotani, hisashi.kawai, satoshi.nakamura}@nict.go.jp

## Abstract

Korean is an agglutinative and highly inflective language with a severe phonological phenomenon and coarticulation effects, making the development of a large-vocabulary continuous speech recognition system (LVCSR) difficult. Choosing a Korean orthographic word-phrase (*eojeol*) as a basic recognition unit leads to high out-of-vocabulary (OOV) rates, whereas choosing an orthographic syllable (*eumjeol*) unit results in high acoustic confusability. To overcome these difficulties, we propose to construct the speech recognition task as a serial architecture composed of two independent parts. The first part is to perform a standard hidden Markov model (HMM)-based recognition of phonemic syllable units of the actual pronunciation (surface forms). In this way, one phonemic syllable corresponds to one possible pronunciation only. Thus, the lexicon dictionary and OOV rates can be kept small, while avoiding high acoustic confusability. Here, the Korean orthography of written transcription are not yet considered. In the second part, the system then transforms the phonemic syllable surface forms into the desirable orthography of a recognition unit, e.g., *eumjeol* or *eojeol*. To solve this task, a noisy-channel model is utilized, wherein the sequence of phonemic syllables is considered as “noisy” string, and the goal is to recover the “clean” string of Korean orthography. The entire process requires no linguistic knowledge, only annotated texts. The experiments were conducted on a Korean dictation database, where the best system could achieve 91.21% *eumjeol* accuracy and 71.30% *eojeol* accuracy.

**Index Terms:** large-vocabulary continuous speech recognition, Korean language, noisy-channel approach.

## 1. Introduction

Most state-of-the-art large-vocabulary continuous speech recognition (LVCSR) systems typically choose words as the basis for recognition units. This is mainly because the words are long enough to differ from each other in a sufficient number of phonemes, while short enough to be able to cover most material with a reasonable number of frequently occurring word forms [1]. This is basic for Indo-European languages (e.g. English), since the number of word forms is relatively small, and the boundaries between adjacent words are clearly separated by a white space. However, this choice becomes problematic in agglutinative and highly inflective languages like Korean.

In Korean orthography, there are two main components; *eumjeol* and *eojeol*. (1) An *eumjeol* is a basic unit that represents an orthographic syllable of a single *Hangul* character. It is basically composed of one to three *jamo* (orthographic phoneme segments); (2) An *eojeol* is a sequence of one or more *eumjeol*, separated by spaces [2]. Although a space exists in the Korean writing script, the word boundary is difficult to define without

any morphological analysis. Because an *eojeol* is a long unit of an agglomerate of morphemes, it is semantically similar to phrases (two or three words) in English. Due to the agglutinative process it may combine one or several stem morphemes with one or several functional morphemes (e.g., tenses, suffixes, honorifics). Consequently, there may be thousands of distinct *eojeol* that can be generated from a given word root, depending on their usage. Thus choosing *eojeol* as a basic recognition unit leads to an extremely high language model perplexity and a large number of out-of-vocabulary (OOV) rates. On the other hand, choosing *eumjeol* as a basic recognition unit may provide small dictionary sizes and OOV rates. However, due to a severe phonological phenomenon in the Korean language, the same orthographic transcription can have a large distinct surface pronunciations depending on the neighboring morphemic and phonemic contexts. Thus, acoustic confusability of *eumjeol* is also increased significantly resulting in low performance of LVCSR.

Many Korean LVCSR systems existing today attempted to overcome these difficulties by creating a set of new units that lie between these two *eojeol* and *eumjeol* units. One way is to choose a morpheme as a basic recognition unit, which has often been used in many agglutinative languages [3, 4]. One study [5] shows that this morpheme-based approach still requires an additional cross-word phone variation lexicon to deal with a severe coarticulation problem. Another study [6] then proposed merging several morphemes into a basic unit and defining it as a word. Starting from the original morpheme units defined in Korean morphology, pairs of short and frequent morphemes are merged into larger units by combining both rule-based and statistical methods. However, this requires a great deal of effort to develop a morphological analysis that involves a great deal of linguistic knowledge about the morphological structure. Another way is to determine appropriate vocabulary units using a data-driven approach [7].

In this paper, we investigate another approach, wherein we perform a standard hidden Markov model (HMM)-based recognition of phonemic syllable units of the actual pronunciation (surface forms). In this way, one phonemic syllable symbol has only one phonetic transcription that corresponds to one possible pronunciation. Thus, the lexicon dictionary and OOV rates can be kept small, while avoiding high confusability due to a severe coarticulation problem. Here, the Korean orthography of written transcription are not yet considered. In the second step, the system then transforms the phonemic syllable surface forms into the desirable orthography of a recognition unit, e.g., *eumjeol* or *eojeol*. To solve this task, a noisy-channel model is utilized, wherein the sequence of phonemic syllables is considered as “noisy” string, and the goal is to recover the “clean” string of Korean orthography.

In the next section, we give an overview of phonological

changes in the Korean language. Then, we describe the LVCSR framework in Section 3, and the noisy-channel approach in Section 4. Details on the experiments are presented in Section 5 and conclusions are drawn in Section 6.

## 2. Phonological Changes in Korean

The orthographic transcription is converted into phonetic surface forms via a phonological process. Phonological changes can occur between within a morpheme and across adjacent morphemes. These changes include consonant and vowel assimilation, dissimilation, insertion, deletion, and contraction [8]. Some examples of the problematic sets of (completely or partially) orthographic *eojeol* are describes as follows:

- An obstruent syllable coda /g/ is nasalized when followed by an syllable onset nasal /m/:  
 Meaning: “citizens”  
 Eojeol (romanized): **/gug-min/**  
 Eumjeol (romanized): **/gug/ /min/**  
 Phonemic syllable: **/gung/ /min/**  
 Phonemes: **/g/ /u/ /ng/ /m/ /i/ /n/**
- A syllable onset lenis obstruent /b/ after syllable coda /h/ is fused with the /h/ to result in an open syllable followed by an aspirated obstruent syllable onset:  
 Meaning: “study of law”  
 Eojeol (romanized): **/beob-hag/**  
 Eumjeol (romanized): **/beob/ /hag/**  
 Phonemic syllable: **/beo/ /pag/**  
 Phonemes: **/b/ /eo/ /p/ /a/ /g/**

The complete phonological phenomenon is rather complicated. A detailed discussion of Korean phonological phenomena can be found in [2, 9].

## 3. Speech Recognition Framework

Given the feature vectors  $x = [x_1, x_2, \dots, x_T]$  of the speech signals, the state-of-the-art statistical speech recognition task is to find an orthographic Korean sequence  $w_e = [w_{e1}, w_{e2}, \dots, w_{eN}]$  that maximizes the conditional probability  $P(w_e|x)$ :

$$\hat{w}_e = \arg \max_{w_e} P(w_e|x). \quad (1)$$

By introducing an intermediate symbol of the phonemic syllable surface form  $s_p = [s_{p1}, s_{p2}, \dots, s_{pM}]$ , the above equation becomes:

$$\begin{aligned} \hat{w}_e &= \arg \max_{w_e} \left\{ \sum_{s_p} P(w_e, s_p|x) \right\} \\ &\approx \arg \max_{w_e} \left\{ \max_{s_p} P(w_e|s_p) P(s_p|x) \right\} \\ &\approx \arg \max_{w_e} P(w_e|\hat{s}_p); \text{ where: } \hat{s}_p = \arg \max_{s_p} P(s_p|x). \end{aligned} \quad (2)$$

This equation suggests that the speech recognition task can be constructed as a serial architecture composed of two independent parts:

- (1) The first part represents finding the most probable phonemic syllable sequence  $\hat{s}_p$ . It is performed by standard HMM-based speech recognition where a phonemic syllable unit is used as the recognition unit as follows:

$$\hat{s}_p = \arg \max_{s_p} P(s_p|x) = \arg \max_{s_p} P(x|s_p)P(s_p), \quad (3)$$

$P(s_p)$  denotes a *language model (LM)* of phonemic syllable units and  $P(x|s_p)$  denotes an *acoustic model (AM)*.

- (2) The second part represents finding the most probable orthographic  $w_e$  sequence given the phonemic syllable sequence  $\hat{s}_p$ . Here, this orthographic  $w_e$  sequence can be either an *eumjeol* or *eojeol* sequence. It is performed by a noisy-channel model as described in the following section.

## 4. Noisy-Channel Approach

The noisy-channel approach utilized here is adopted from statistical machine translation (SMT). Following the SMT terminology, we define the phonemic syllable  $s_p$  sequence to be the source language that has to be translated into the target language, namely the orthographic  $w_e$  sequence. SMT formulates the translation process as the maximization problem of the conditional probability:

$$\hat{w}_e = \arg \max_{w_e} P(w_e|s_p) = \arg \max_{w_e} P(s_p|w_e)P(w_e), \quad (4)$$

where  $P(w_e)$  denotes an LM representing the likelihood of the target language of the orthographic unit  $w_e$  and  $P(s_p|w_e)$  denotes a *translation model* representing the generation probability from phonemic syllable  $s_p$  into orthographic unit  $w_e$  [10]. During the translation process (*decoding*), a score based on the statistical model probabilities is assigned to each translation hypothesis and the one that gives the highest probability is selected as the best translation. The basic framework of SMT is illustrated in Fig. 1.

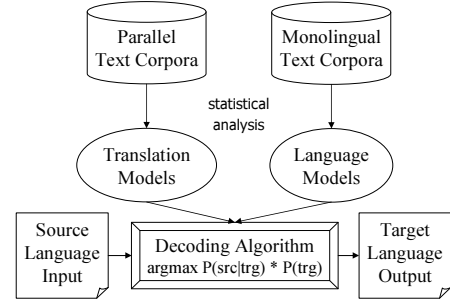


Figure 1: Basics of SMT Framework.

In practice, we applied our existing SMT system without any modification. This is based on phrase-based machine translation techniques [11] integrating within a log-linear framework [12]. For the training of the SMT models, word alignment [13] and language modeling [14] toolkits were used. N-gram language models built with Knesser-Ney smoothing [15] were used along with a lexicalized distortion model [16]. The system is trained in a standard manner, using a minimum error-rate training (MERT) procedure [17] with respect to the BLEU score [18] on held-out development data to optimize the log-linear model weights. For decoding, a multi-stack phrase-based SMT decoder called CleopATRA [19] was used.

Assuming one phonemic syllable unit is one “word” of source language unit, then a “phrase” is one segment of the source phonemic syllables sequence. The basic idea is to segment the source phonemic syllables sequence into subsequences (phrases), then translate each phrase individually, and finally compose the target orthographic unit sequence based on the n-gram language model. As the process is phrase-based, the translation is generated phrase-by-phrase, and the orthographic unit  $w_e$  sequence of the target language is typically generated in order in a forward manner. In this way, it allows the transformation of several phonemic syllables into one orthographic unit  $w_e$  (e.g., a single *eojeol*). More details of our SMT system can be found in [19].

## 5. Experimental Evaluation

### 5.1. Data Corpora

The experiments were conducted using the large vocabulary continuous Korean speech database developed by the Speech Information Technology and Industry Promotion Center (SiTEC) [20]. The three types of data corpora used here are referred to as Sent01, Dict01 and Dict02:

- Phonetically-balanced sentences (Sent01)  
About 20,000 sentences selected from a large Korean text corpus on the basis of containing morphemes of high frequency, including phonetically balanced sentences, were used and organized into 200 prompt sets. The clean speech of 200 speakers (100 males, 100 females) was recorded in a soundproof room (each speaker uttered one prompt set of about 100 sentences).
- Dictation application sentences (Dict01 and Dict02)  
About 40,000 sentences composed on the basis of containing words and morphemes of high frequency were used and organized into 200 prompt sets. This data was generated for a dictation application. The clean speech of 400 speakers (200 males, 200 females) for each Dict01 and Dict02 was recorded in a soundproof room (each speaker uttered about 100 sentences).

The romanized transcriptions of both orthographic and surface form were provided for the whole corpus. The last two prompt sets from each Sent01, Dict01 and Dict02 were allocated to the test set, while the remaining data were used as the training set. The *eojeol* and *eumjeol* coverage on training and test set of SiTEC corpora are described in Table 1.

Table 1: *Eojeol* and *eumjeol* coverage in SiTEC corpora

	Training	Test
# Sentences	101,714	1,238
# Eojeols	863,556	11,119
# Eumjeols	2,275,899	27,842
# Unique Eumjeol	1,177	613
# Eumjeol Lexical Entries	4,252	-
# Unique Phonemic Syllable	1,337	679
# Phonemic Syllable Lexical Entries	1,337	-

### 5.2. Baseline LVCSR System

Our baseline HMM-based acoustic model was trained using all Sent01, Dict01 and Dict02 training data. A sampling frequency of 16 kHz, frame length of a 20-ms Hamming window, frame shift of 10 ms, and 25 dimensional feature parameters consisting of 12-order MFCC,  $\Delta$  MFCC and  $\Delta$  log power were used as feature parameters.

The full phoneme set, as defined in [21], contained a total of 40 phoneme symbols. These consisted of 19 consonants and 21 vowels (including nine monophthongs and 12 diphthongs). One silence symbol was added during acoustic model training. Three states were used as the initial model for each phoneme. Then, a state level HMnet was obtained using a successive state splitting (SSS) algorithm based on the minimum description length (MDL) criterion in order to gain the optimal structure in which triphone contexts are shared and tied at the state level. Details about MDL-SSS can be found in [22]. The resulting context-dependent triphone had 2,231 states in total with 5, 10, 15 and 20 Gaussian mixture components per state.

The pronunciation dictionary was composed based on orthographic syllable (*eumjeol*) units with multiple pronunciations. The *eumjeol* bigram and trigram language models were

trained using the text data of Sent01, Dict01, and Dict02, yielding a trigram perplexity of 16.6 on Sent01, 20.6 on Dict01, and 31.2 on the Dict02 test set with an OOV-rate of less than 1%.

The performance of the baseline system for each Sent01, Dict01, and Dict02 test sets is shown in Fig. 2. The best models only achieved 70.13%, 67.96%, and 57.39% syllable accuracy on Sent01, Dict01, Dict02, respectively. Note that this performance can also be considered as character accuracy to fairly compare with other languages. As can be seen, even with a small dictionary and low OOV rates, orthographic syllable recognition is still difficult due to high acoustic confusability.

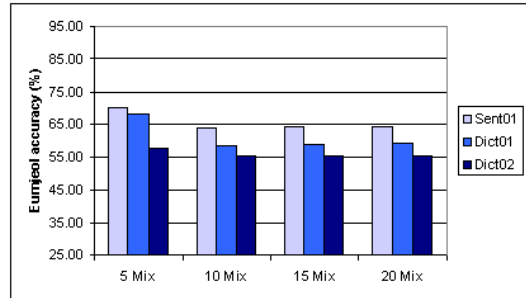


Figure 2: Orthographic syllable (*eumjeol*) accuracy of baseline LVCSR system.

### 5.3. Proposed LVCSR System

As described in Section 3, we constructed our system as a serial architecture of two independent parts: (1) finding the most probable phonemic syllable sequence  $\hat{s}_p$ ; and (2) translates the phonemic syllable  $\hat{s}_p$  sequence into an orthographic  $\hat{w}_e$  sequence.

In the first step, phonemic syllable recognition was performed using the same training set and acoustic model as in the baseline LVCSR. The only differences are the pronunciation dictionary and language model. Instead of using orthographic syllables (*eumjeol*), the lexical entry and language model unit were composed based on the phonemic syllable units (surface forms). Thus no multiple pronunciations exist here, and the resulting lexicon size is only one-third of the baseline lexicon. The language model had a slightly higher trigram perplexity of 18.7 on Sent01, 22.4 on Dict01, and 31.3 on the Dict02 test set with OOV rates of less than 1%. The performance on phonemic syllable accuracy for each Sent01, Dict01, and Dict02 test set is summarized in Table 2. The best model achieved a very high performance of 88.67%, 88.71%, and 80.17% phonemic syllable accuracy on Sent01, Dict01, Dict02, respectively. However, these results can not yet be compared with the baseline performance, since they are based on different unit symbols.

Table 2: Phonemic syllable accuracy of proposed system

Test Set	5 Mix	10 Mix	15 Mix	20 Mix
Sent01	86.99	88.13	88.67	88.48
Dict01	87.75	87.87	88.71	88.33
Dict02	78.50	79.32	80.17	80.42

For the second step, we first trained the noisy-channel model on the same training set. The performance of the proposed noisy-channel model given the correct input transcription of the phonemic syllable sequence is given in Table 3. We then utilized the trained noisy-channel model on the output of phonemic syllable recognition. The performance on both *eojeol* and *eumjeol* target unit sequences for each of the Sent01, Dict01, and Dict02 test sets was shown in Fig. 3 and 4, respectively. Note that this performance can also be considered

for word-phrase and character accuracy to compare fairly with other languages. The best system could achieve 71.30% *eojeol* accuracy and 91.21% *eumjeol* accuracy (on Sent01 test set), yielding 26.67% absolute improvement with respect to baseline orthographic syllable recognition.

Table 3: *Eojeol* and *eumjeol* accuracy provided by noisy-channel model given correct input of phonemic syllable transcription.

Test Set	Eojeols	Eumjeols
Sent01	86.82	98.99
Dict01	94.55	98.50
Dict02	83.47	97.72

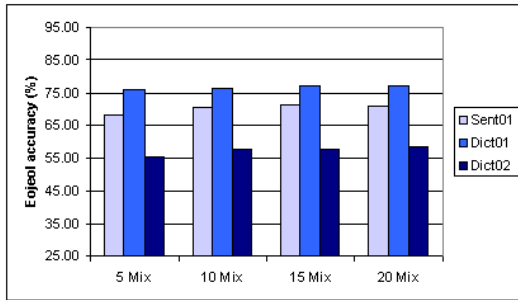


Figure 3: Orthographic word-phrase (ejojeol) accuracy provided by noisy-channel model.

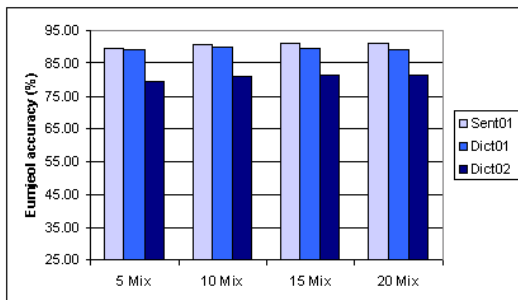


Figure 4: Orthographic syllable (eumjeol) accuracy provided by noisy-channel model.

## 6. Conclusions

We demonstrated the possibility of utilizing a noisy-channel model in a Korean LVCSR that allows transformation between different symbols in Korean orthography. The noisy-channel model is applied after we perform the standard HMM-based recognition on phonemic syllable of the actual pronunciation (surface forms). The goal is to output a sequence of Korean orthography given a sequence of phonemic syllable surface forms. We attempted to transform phonemic syllable surface forms into both orthographic *ejojeol* and *eumjeol*. The best system could achieve 71.30% *ejojeol* accuracy and 91.21% *eumjeol* accuracy. The results reveal that the proposed method can help to improve Korean LVCSR performance, giving 26.67% absolute improvement with respect to baseline orthographic syllable recognition. The entire process requires only annotated texts without any linguistic knowledge, making it applicable to other agglutinative languages.

## 7. Acknowledgements

The authors would like to thank Andrew Finch and Chooi-Ling Goh for their support and useful discussion regarding the SMT framework.

## 8. References

- [1] A. Waibel, P. Geutner, L. Mayfield-Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *IEEE Special Issue on Spoken Language Processing*, vol. 88, no. 8, pp. 1297–1313, 2000.
- [2] K. Yoon and C. Brew, "A linguistically motivated approach to grapheme-to-phoneme conversion for Korean," *Computer Speech & Language*, vol. 20, no. 4, pp. 357–381, 2006.
- [3] G. Choueiter, D. Povey, S. Chen, and G. Zweig, "Morpheme-based language modeling for Arabic LVCSR," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 1053–1056.
- [4] K. Kirchoff and R. Sarikaya, "Processing morphologically-rich languages," in *INTERSPEECH Tutorial*, Antwerp, Belgium, 2007.
- [5] H.-J. Yu, H. Kim, J.-M. Hong, M.-S. Kim, and J.-S. Lee, "Large vocabulary Korean continuous speech recognition using a one-pass algorithm," in *Proc. ICSLP*, Beijing, China, 2000, pp. 278–281.
- [6] O.-W. Kwon and J. Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," *Speech Communication*, vol. 39, pp. 287–300, 2003.
- [7] D. Kiecza, T. Schultz, and A. Waibel, "Data-driven determination of appropriate dictionary units for Korean LVCSR," in *Proc. ICSP*, Seoul, Korea, 1999, pp. 323–327.
- [8] B. Kim, G. Lee, and J. Lee, "Morpheme-based grapheme to phoneme conversion using phonetic patterns and morphophonemic connectivity information," *ACM Transactions on Asian Language Information Processing*, vol. 1, no. 1, p. 6582, 2002.
- [9] S.-C. Song, *201 Korean Verbs - Fully Conjugated in All Forms*. Baron's Educational Series, 1988.
- [10] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.
- [11] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc of the Human Language Technology Conference*, 2003, pp. 127–133.
- [12] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. of ACL*, Philadelphia, PA, USA, 2002, p. 295302.
- [13] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.
- [14] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. of ICSLP*, Denver, USA, 2002, pp. 901–904.
- [15] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. ICASSP*, 1995, pp. 181–184.
- [16] Y. Al-Onaizan and K. Papineni, "Distortion models for statistical machine translation," in *Proc. ACL/COLING*, Sydney, Australia, 2006, pp. 529–536.
- [17] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. of ACL*, Sapporo, Japan, 2003, p. 160167.
- [18] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of ACL*, Philadelphia, USA, 2002, pp. 311–318.
- [19] A. Finch, E. Denoual, H. Okuma, M. Paul, H. Yamamoto, K. Yasuda, R. Zhang, and E. Sumita, "The NICT/ATR speech translation system for IWSLT 2007," in *Proc. IWSLT*, Trento, Italy, 2007, pp. 103–110.
- [20] B. Kim, D. Choi, Y. Kim, K. Lee, and Y. Lee, "Current state and future plants at SITEC for speech corpora for common use," *Malsori*, vol. 46, p. 175186, 2003.
- [21] M. Kim, Y. Oh, and H. Kim, "Non-native pronunciation variation modeling using an indirect data driven method," in *Proc. ASRU*, Kyoto, Japan, 2007, pp. 231–236.
- [22] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 8, pp. 2121–2129, 2004.