Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR

Daniel Kiecza, Tanja Schultz and Alex Waibel

Interactive Systems Laboratories
University of Karlsruhe (Germany), Carnegie Mellon University (USA)
{*kiecza,tanja,waibel*}@*ira.uka.de*

ABSTRACT

This paper describes the design of our Korean large vocabulary speech recognition system using the multilingual dictation database GlobalPhone. Defining appropriate dictionary units for this purpose is not a trivial task since using word phrases (eojeols) gives very high OOV-rates, above 30%, whereas using syllable units results in high confusabilities and a very limited scope of standard language models. We investigate a data-driven approach which overcomes these limitations. The results show that the data-driven approach reduces the OOV-rate to below 1% and significantly outperforms the syllable based approach according to phone and syllable accuracy giving 79.4% and 69.3% accuracy respectively. For our best system we present lattice based accuracies achieving 95.0% syllable accuracy and 82.7% eojeol accuracy.

1. Introduction

Korean is an inflected language, i.e. words are composed by concatenating one to several particles to the word stem in order to indicate mode and tense of verbs or case, number, and gender of nouns. Therefore the choice of appropriate dictionary and language model[1] units for an HMM based Korean LVCSR system is difficult. Using the compound units (eojeols) that result from the agglutination process as dictionary units gives unmanageably large dictionaries with extremely high Out-of-Vocabulary (OOV) rates [2].

Korean words are built from only about 3500 different syllables, where each syllable consists of one to four phonemes. Choosing these syllables as dictionary units provides small dictionaries and OOV-rates far below one percent. Unfortunately, due to their shortness, two problems arise when this approach is used:

- acoustic confusability of syllable units is increased,
- a standard trigram language model has very limited scope using these short units.

We present a data-driven method that attempts to overcome the difficulties of using either eojeols or syllables as units by creating a set of units that lie "between" these two extremes. The basic idea is to start from the syllable based system and to repeatedly merge units in order to decrease their acoustic confusability.

To evaluate our approach, several recognizers are trained and tested using different unit sets. For all our experiments we used the GlobalPhone dictation database which currently consists of 15 languages [3, 4].

[1] We use the same set of units for dictionary and language model. Dictionary units means dictionary and language model units throughout this paper.

2. Databases

2.1. Acoustic Database

For development and evaluation of our systems we use the Korean portion of the GlobalPhone database [3, 4]. This section consists of 20 hours of speech data spoken by 100 native Korean speakers. Every speaker read several articles from a Korean national newspaper. The articles were chosen from the areas: national politics, international politics, and economy. The speech data was recorded at a sampling rate of 48kHz using a close-talking microphone connected to a DAT-recorder.

	Train	Test
Speakers	80	10
Utterances	6,350	84
Vocabulary (eojeol)	41,876	923
OOV words	–	41.43% (440)
Total utterances	6,434	
Total vocabulary (eojeols)	42,310	

Table 1: Summary of acoustic database.

After transferring the sound data from DAT to hard disc it was downsampled to 16kHz, 16-bit. Eighty of the speakers were used for training of the acoustic models, ten were defined as test set. The remaining ten are kept as a further cross-validation set. A subset of 84 uniformally selected utterances from the test set was used to carry out our experiments. See table 1 for an overview of the database.

2.2. Language Model Data

To overcome the sparse data problem in language model generation we collected a large corpus of text data from the internet. The online newspaper articles of the Korean newspaper *Chosunilbo* can be retrieved from the URL *http://www.chosun.com/w21data/html/news/*. We used the Unix tool *wget* to get all articles from October 1995 to August 1998. A text preprocessing script cleaned the text data by removing all HTML-related code. Numbers were mapped onto their textual transcription. Acronyms were replaced by mapping each letter onto its pronunciation. The script then dropped all sentences which still contained non-hangul characters as our speech recognition system is based on a pure hangul database.

The resulting text corpus has a total size of 15,413,927 eojeols. It consists of 1,484,557 different eojeols. In terms of syllables the total corpus size is 43,764,433 and the vocabulary size is 3,578. To ensure a time-efficent evaluation of our unit determination process we decided to use only a part of this corpus (about 15%) and to keep the rest for future fine-tuning of the systems.

This text corpus portion has a total size of 2,261,773 eojeols. It consists of 400,400 different eojeols. In terms of syllables the corpus size is 6,551,344 and the vocabulary size is 2,980, see table 2. This portion plus the transcription data of the training utterances were used together as a basis for our merging algorithm as well as for language model generation. In the following we will refer to this data as *chosun+train*.

	All	Chosun+train
Number of eojeols	15,413,927	2,261,773
Eojeol vocabulary size	1,484,557	400,400
Number of syllables	43,764,433	6,551,344
Syllable vocabulary size	3,578	2,980

Table 2: Summary of language model corpus.

3. Dictionary Unit Generation

3.1. Pronunciation Generation

Our recognition systems are based on a romanized form of Korean characters. We transform hangul characters automatically into a romanized transcription using the code conversion tool *hcode* [9].

In order to create the pronunciation dictionary which is needed for our speech recognition system we compiled the phonological rules (like assimilation, reinforcement and weakening) described in [5, 6]. This set of rules was then applied to each corpus word to transform the romanized written form of this word into a sequence of corresponding phones.

Handling phonological changes inside a unit is straightforward, simply apply the defined set of rules. However, phonological effects can also occur at unit boundaries. To handle these cases we extract the last syllable of the preceding unit and the first syllable of the succeeding unit and connect them respectively to the beginning and end of the current unit.

Now the set of rules can be easily applied, phonological changes happen within the newly created meta-unit. After the corresponding sequence of phones is created the phones that belong to the two added syllables are removed. As a result we obtain the pronunciation of the current unit in the given context.

Of course this procedure might return different pronunciations for a specific unit depending on the context. These are handled as pronunciation variants in the recognizer's dictionary.

3.2. Merging Concept

Our goal is to generate a set of dictionary units which on one hand are longer than syllables, reducing acoustic confusability, and increasing the range of the trigram language model. On the other hand the units must be shorter than eojeols so that OOV rate is manageably low.

A lot of human knowledge and expert effort is required to build morphological tagging systems which can be used to generate appropriate dictionary units for Korean speech recognition [1]. Instead we use a *data-driven*, statistical approach that requires no a-priori linguistic knowledge.

The starting point for our unit determination approach is the syllable based recognition system. We repeatedly merge units to form new longer units[2] until a stop criterion is reached.

As a preprocessing step to our merging algorithm we first retrieve all syllable pairs that appear in the corpus *chosun+train*. For each syllable pair we generate the pronunciation from center vowel to center vowel, for example han-kuk $\rightarrow$ A N K U. Pronunciation generation is done automatically as described in section 3.1.

The general merging process is controlled by the following data-driven iterative procedure:

1. Choose a pronunciation transition according to a specified rule and/or select the syllable pair(s) that produce this transition according to another specified rule.
2. Merge all these syllable pair(s) in the corpus *chosun+train*.

[2] We tag units that are not at the beginning of an eojeol with a preceding dash. This way it is straightforward to extract eojeols from a syllable or merge based hypothesis.

One can think of different stop criteria for this algorithm, e.g. perplexity based or OOV rate based. We chose an arbitrary OOV rate of 5% as the stop criterion. The recognition toolkit used for the evaluation process can handle a maximum of 64k words in the recognition vocabulary. However, the OOV rate is still below 1% for the resulting merged systems when the maximum vocabulary has been reached. We evaluated two systems with different merging approaches: *MergeAll* and *MergeMax*. They work as follows:

MergeAll Find the one or more pronunciation transition with the highest frequency in the text corpus. Merge all syllable pairs that produce this pronunciation transition in the corpus.

MergeMax Find the one or more pronunciation transition with the highest frequency in the text corpus. Merge the syllable pair that produces this pronunciation transition the most often. Thus *MergeMax* can be considered as a more selective variation of *MergeAll*.

Figure 1 shows a logarithmic self coverage diagram of the corpus *chosun+train* for four different dictionary unit sets: Syllable, MergeAll, MergeMax and Eojeol.

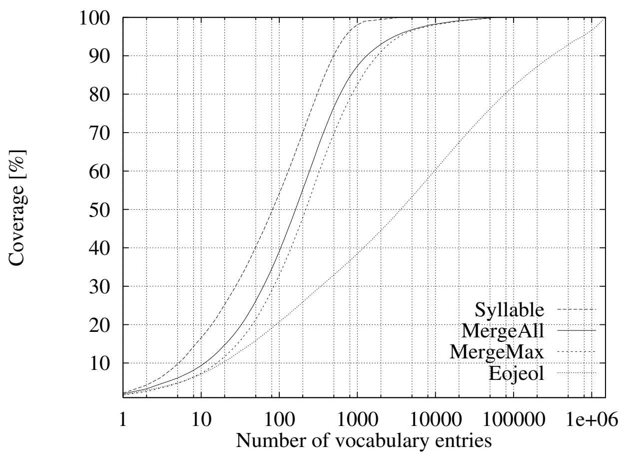

Figure 1: Self coverage of language model corpus *train+chosun*.

The cross coverage of the test corpus using *train+chosun* is displayed in figure 2. At a vocabulary size of 1.25 million eojeols a maximum cross coverage of 88% is reached. Using the 64k most frequent eojeols yields a cross coverage of 69%.

4. Experiments

4.1. The Janus Speech Recognition Toolkit

The recognition results presented in this paper were obtained using the Janus Recognition Toolkit (JRTk) [7, 8]. We defined a set of 48 phones. Each of them is modeled by a three state, left-to-right HMM with 16 diagonal Gaussian mixtures per state. The preprocessing consists of extracting Mel-frequency cepstral coefficents every 10ms with a window size of 20ms. The final 24 dimensional feature vector is computed by a truncated LDA transformation of the 41 dimensional vector consisting of the 13 MFCCs, their first and second order derivatives, the energy value and zero crossing. Vocal tract length normalization and cepstral mean subtraction are used to minimize speaker and channel differences.

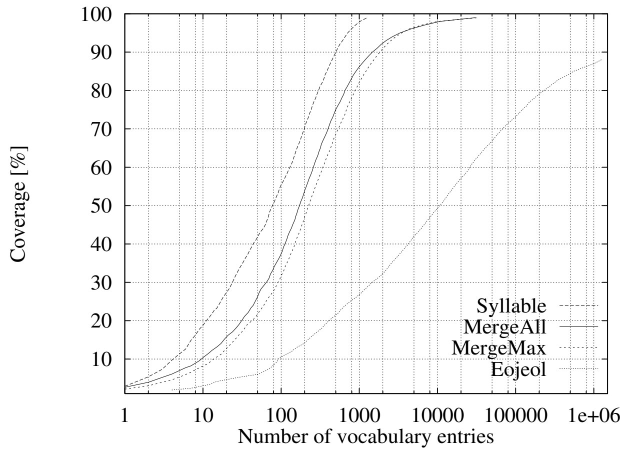

Figure 2: Cross coverage of language model corpus *train+chosun* and test corpus.

An initial context-independent Korean recognition system was trained using the labels generated by a speaker-adapted multi-lingual (German, English, Japanese, and Spanish) recognizer. The Korean phones were initialized by their closest multi-lingual equivalents. All context-dependent systems consisted of 3000 quintphone models. The decision tree used for these models was generated using a set of 63 phone context questions.

4.2. Phone Set

Based on [5, 6] we defined a total of 48 phones, 9 vowels, 11 diphthongs and 28 consonants. Furthermore, we have one silence model and an acoustic model that represents human non-speech noises.

This phone set is very detailed and consequently a few models were rather poorly estimated. Therefore we reduced the number of phones to 41 for further experiments. The three poorly estimated diphthongs /o-ɛ/, /i-ɛ/, /u-e/[3] are split up into their monophthongs. Each of the four consonants ch, p, t and k[4] are represented by only one phone model instead of two.

[3] IPA-symbols

[4] McCune-Reischauer transcription symbols

4.3. Results

The recognition accuracy results are summarized in table 3 and show that the merged systems improved recognition performance for both syllable and phone recognition accuracy.

	Syllable	Phone
Syllable	62.8	74.3
MergeMax	68.8	79.1
MergeAll	69.3	79.4

Table 3: Recognition accuracy values, %.

From table 4 we can see that the baseline syllable system has the smallest perplexity value of the three systems because it has a vocabulary size of only 2,980 units whereas the vocabulary size of the merged systems is as big as 64k.

	OOV	Perplexity	Normalized PP
Syllable	0.016	37.7	37.7
MergeMax	0.700	129.9	43.9
MergeAll	0.695	90.9	43.2

Table 4: Language model characteristics.

The merging approach creates new, longer units initialized using the syllable system. Merging units can create new vocabulary entries that only occur in the test set of our data but not in *train+chosun*. In this case OOV can only increase during the merging process. But, although OOV is higher for the merged systems it remains below 1%.

The merged systems have an average unit length of 1.97 and 1.88 syllables (MergeMax and MergeAll respectively). These longer units increase our polyphone modelling potential as can be seen in figure 3. The average number of units for a pronunciation sequence in the recognizer's dictionary is smaller for the merged systems than for the syllable systems, see table 5. Thus while the task complexity of the merged systems is higher than the task complexity of the syllable system we have longer and less confusable units while retaining low OOV.

We measured the systems' accuracy using three different criteria, eojeol accuracy, syllable accuracy and phone accuracy. Table 3 shows significant improvement in phone and syllable recognition performance using the merged systems over the baseline approach.

Although eojeol accuracy did increase with our merged systems it didn't increase as significantly as the phone and syllable accuracy. This is because the average length of an eojeol in *train+chosun* is 2.91 syllables, so a trigram language model built on our merged units can on average reach only into the next eojeol unit. This language model is still not powerful enough to perform well on eojeol bases.

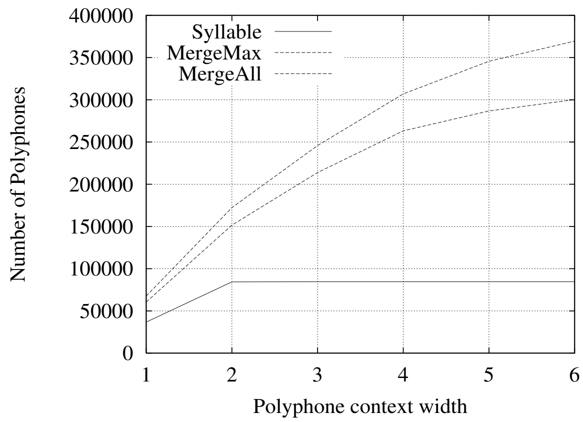

Figure 3: Number of polyphones in training corpus.

We also measured the overall lattice word accuracy for each system. The lattice word accuracy (LWA) is the word accuracy of that path in the word hypothesis graph that comes closest to the reference sentence. Thus it defines an upper bound for the word accuracy we can get from a lattice. Table 6 shows the lattice word accuracy for the three systems. The syllable system outperforms the two merged systems although it was worse in performance with respect to phone and syllable accuracy.

	Average number of units
Syllable	2.35
MergeMax	1.65
MergeAll	1.17

Table 5: Average number of units by which a pronunciation sequence in the dictionary is produced.

	Syllable	Merge Unit	Eojeol
Syllable	93.1	–	75.1
MergeMax	87.5	90.0	70.3
MergeAll	88.1	90.5	70.2

Table 6: Lattice word accuracy, %.

These results are surprising compared to the results in table 3 because the merged systems did not perform as well as the syllable system. We analysed the LWA hypothesis results and found for about 30% of the test utterances errors that resulted from deletions at either end of a sentence, especially the beginning. One explanation for this phenomenon could be that the speaker utterances have been segmented too sharply. This of course makes it very difficult for the first reference word to be recognized properly. For the syllable based recognizer this means at least one misrecognized syllable. But for a merged system this means at least one misrecognized merged unit which consists of almost two syllables on an average. As a consequence a merged system performs worse than a syllable system when comparing their syllable LWA.

Using a morphological approach Kwon et al. [1] achieved a syllable (character) accuracy of 90.8% and an eojeol accuracy of 81.3%. Their most similar system to our best syllable system achieved 84.5% syllable accuracy and 69.6% eojeol accuracy. But results can't be compared directly as Kwon et al. used a different task for their experiments.

4.4. System Improvements

We evaluated further improvements to the syllable system. Firstly, we applied the phone set reduction discussed in section 4.2 to ensure reliable estimation of all phone models. Secondly, we introduced a new phone context question to our cluster algorithm. This question is about whether the current phone is on a merge unit boundary.

Together these improvements gave a relative phone accuracy improvement of 16.3% and a relative syllable accuracy improvement of 13.7%. This results in a phone recognition accuracy of 78.5% and a syllable recognition accuracy of 67.9%. The syllable LWA of this system is 95.0%, the eojeol LWA is 82.7%. These results are summarized in table 7.

	Syllable	Phone
Accuracy	67.9	78.5
Relative improvement	13.7	16.3
	Syllable	Eojeol
Lattice accuracy	95.0	82.7
Relative improvement	27.5	30.5

Table 7: Improved system results, %.

5. Conclusion

In this paper we presented a new approach to generate dictionary units for Korean LVCSR systems. Unlike a morpheme based recognition system this approach does not use human knowledge but is completely *data-driven*. We achieved 79.4% phone recognition accuracy and 69.3% syllable recognition accuracy. Lattice based accuracies were 95.0% for the syllable case and 82.7% for the eojeol case.

Future work will be focused on an implementation of a more sophisticated language model which operates on word hypothesis graphs (lattices). We will verify the LWA results by adding several frames to the beginning and the end of each utterance and carrying out the experiments once again. Furthermore we will build a recognition system based on morpheme units to compare the different approaches more closely.

Acknowledgement

The authors wish to thank all members of the Interactive Systems Labs, especially Michael Finke for many fruitful discussions and Iain Matthews for proof-reading the paper. Many thanks to the Korean GlobalPhone team Sang-Hun Shin, Keal-Chun Cho and Kyung-Kyu Lee. This research would not have been possible without their great enthusiasm during collection and validation of the database.

References

1. Kwon, Oh-Wook and Hwang, Kyuwoong and Park, Jun: *Korean Large Vocabulary Continuous Speech Recognition Using Pseudomorpheme Units* to appear in: Proc. Eurospeech 1999, Budapest, 5.–9. September
2. Lee, Hang-Seop and Park, Jun and Kim, Hoi-Rin: *An Implementation of Korean Spontaneous Speech Recognition System* in: Proc. ICSPAT96, pp. 1801–1805, Seoul, Korea, 1996
3. Schultz, Tanja et al.: *Language Independent and Language Adaptive Large Vocabulary Speech Recognition* in: Proc. ICSLP, pp. 1819–1822, Sydney, Australia 1998
4. Schultz, Tanja et al.: *The GlobalPhone Project: Multilingual LVCSR with Janus-3* in: Proc. SQEL, pp. 20–27, Plzeň 1997.
5. Herrmann, Wilfried: *Lehrbuch der modernen koreanischen Sprache*, Helmut Buske Verlag, Hamburg, 1994
6. Seok Choong Song: *201 Korean Verbs - fully conjugated in all the forms*, Barron's Educational Series, Inc., 1988
7. Lavie, Alon and Waibel, Alex and Levin, Lori and Finke, Michael and Gates, Donna and Garvalda, Marsal and Zeppenfeld, Torsten and Puming, Zhan: *Janus III: Speech-to-Speech Translation in Multiple Languages*, Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing, Munich, Germany, 1997
8. Finke, Michael and Fritsch, Jürgen and Geutner, Petra and Ries, Klaus and Waibel, Alex: *The JanusRTk Switchboard/Callhome 1997 Evaluation System*, Proceedings of the LVCSR Hub5-e Workshop, Baltimore, Maryland, May 1997
9. *http://pantheon.yale.edu/~jshin/faq/qa8.html*: Contains a description of the hangul code translation tool *hcode*.