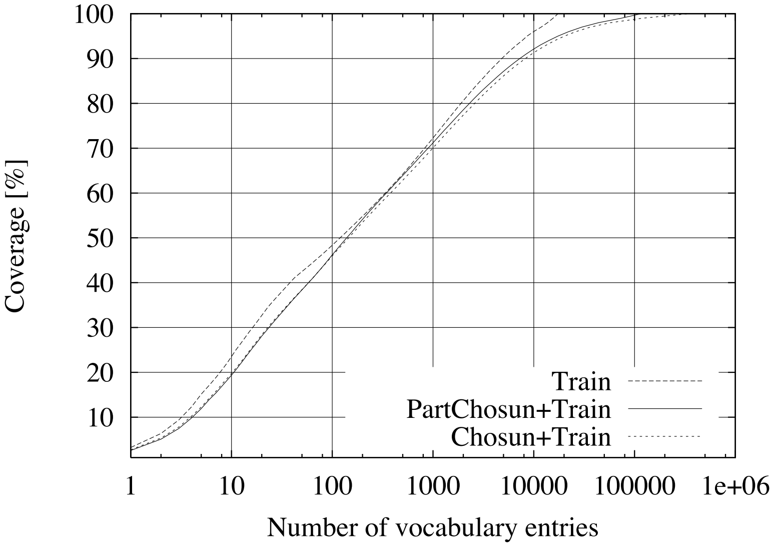

Figure 5.1: Self coverage of *Morph*-based language model corpora.

The analog information for the corpus *Test-Utts* using the same corpora as above is shown in figure D.2 and D.3, respectively. These figures do not differ significantly from the cross coverage figures for *Test* and, as a consequence, are presented in the appendix. We see from these diagrams that the coverage for morpheme units is more consistent than for eojeols; the self coverage and the cross coverage diagrams are almost identical. For the eojeol units the cross coverage is about five percent lower than the corresponding self coverage as can be seen from the respective diagrams in section 4.2.

The *Morph* and *MorphTag* based vocabulary growth is drawn in figure 5.5. We see that the vocabulary grows significantly slower than on an eojeol based corpus. Cross coverage is far above 90% for both *Morph* and *MorphTag* when using a 64k vocabulary. To be precise, the *Morph* system has an OOV rate of 2.37% with a vocabulary of 64k, for *MorphTag* the rate is at 2.78%. These figures are calculated on the corpus *PartChosun+Train*. See table 5.2 for a summary of the OOV information. To summarize, morpheme units are much more suitable for a Korean recognition system than eojeol units.

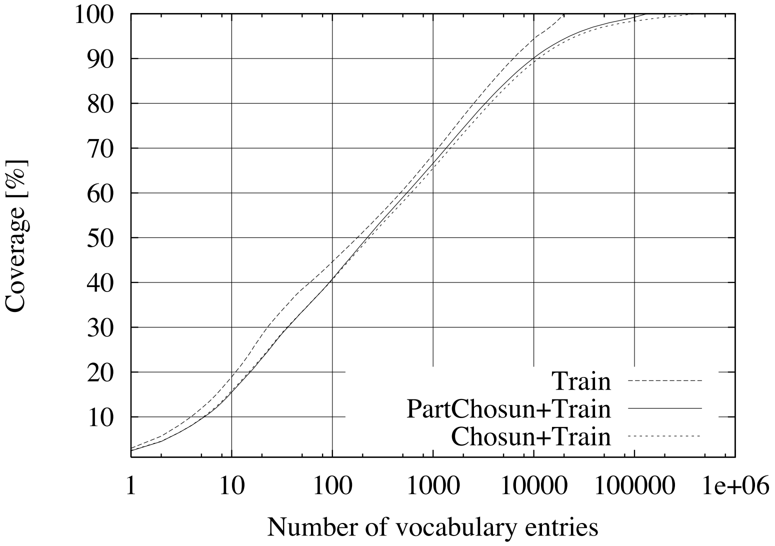

Figure 5.2: Self coverage of *MorphTag*-based language model corpora.

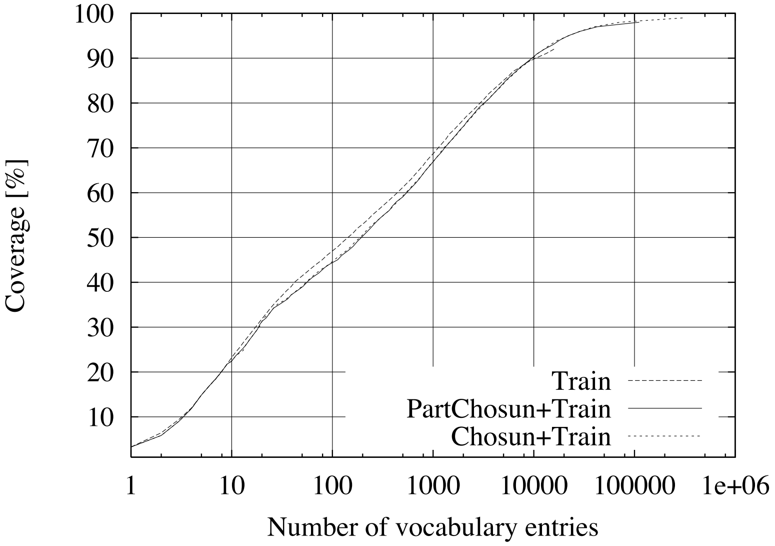

Figure 5.3: *Morph*-based cross coverage of *Test* with different language model corpora.

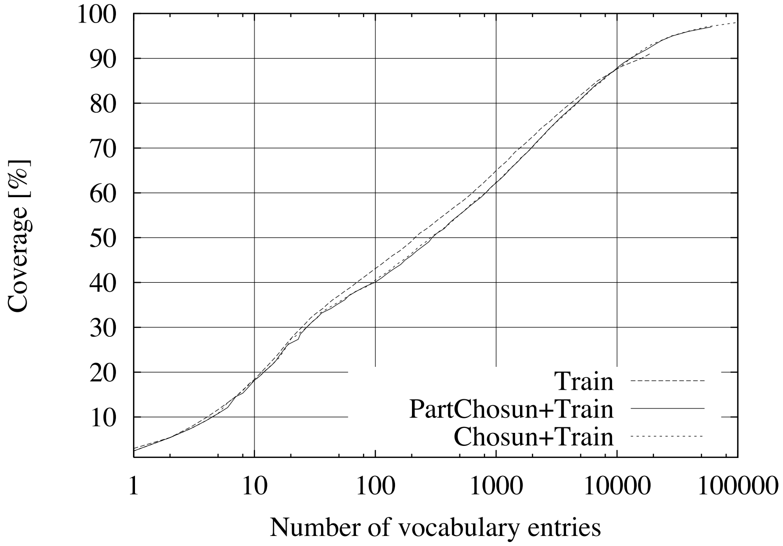

Figure 5.4: *MorphTag*-based cross coverage of *Test* with different language model corpora.

	Test	*Test-Utts*
Morph		
Train	7.44% (1631)	8.40% (176)
PartChosun+Train	1.95% (427)	2.20% (46)
Chosun+Train	0.93% (203)	0.96% (20)
MorphTag		
Train	8.37% (1837)	9.50% (199)
PartChosun+Train	2.22% (486)	2.67% (56)
Chosun+Train	1.12% (245)	1.19% (25)

Table 5.2: Summary of OOV rates and, in parantheses, OOV words.

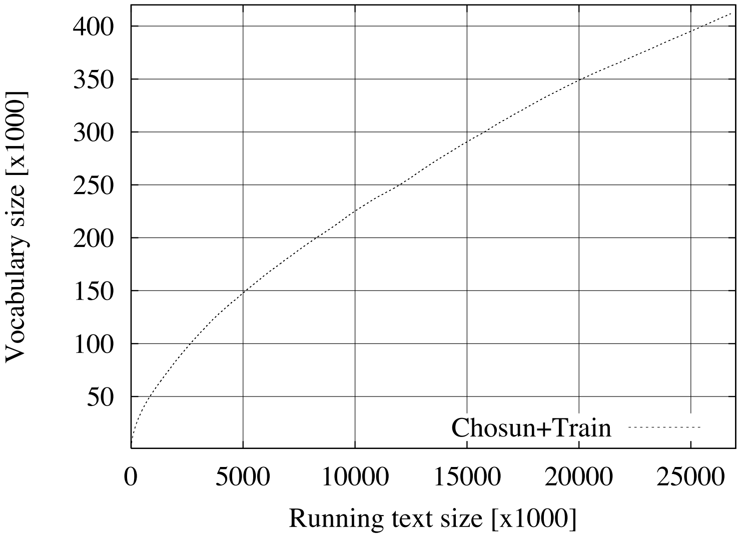

Figure 5.5: Morpheme based vocabulary growth in corpus *Chosun+Train*.

Chapter 6

Data-Driven Unit Determination

6.1 Motivation

Our goal is to generate vocabulary units which on one hand are longer than characters, reducing acoustic confusability, and increasing the range of the 3-gram language model. On the other hand, the units must be shorter than eojeols to retain the OOV rate at a manageably low level.

In preliminary experiments we developed a speech recognition system based on characters as vocabulary units. We presented results of character based recognition systems in [27]. An analysis of the recognition results revealed one dominant type of recognition error: The confusion of pairs of character pairs which share the same sequence of phones from the center vowel of the left character to the center vowel of the right character. We will refer to this phone sequence as *vowel-to-vowel transition*, or simply *transition*. Consider the following examples: The two character pairs *sin eop* and *si neo* share the same vowel-to-vowel transition: *I N EO*. The transition of *math neun* and *manh eun* is *A N EU*, for *ceon hyeo* and *cheo nyeo* it is *EO N iEO*. It is such pairs that are often confused during recognition.

A good language model would certainly help to avoid this type of confusion error. But from an eojeol perspective, the 3-gram language model based on character units has a very limited scope. And in fact, experiments show that this language model does not help to avoid these errors.

The data-driven approach taken here tries to address this problem. Decreasing the average number of confusable vowel-to-vowel transitions between vocabulary units should result in a significant increase in recognition performance. The basic idea of this approach is to repeatedly merge specific unit pairs and thus to *lock* the vowel-to-vowel transition between these pairs *inside* the newly created units. This reduces the number of unit pairs that share a particular vowel-to-vowel transition. In addition the increased average length of vocabulary units should also help the 3-gram language model by increasing its scope.

6.2 Determination of Units

6.2.1 Preprocessing

For the further explanation we consider the sentence *o-neul han-kuk-e ka-yo*[1] as an example.

First, we retrieve all character pairs that appear in the text corpus. The example sentence contains the character pairs *o neul*, *neul han*, *han kuk*, *kuk e*, *e ka* and *ka yo*. Then, for each character pair its vowel-to-vowel transition is generated, for example *han kuk* → A N G U. Pronunciation generation is done automatically as described in section 4.3.

For each such character pair we count how often it appears in the text corpus. A hash table with two keys is used to store this information. The first key in the table is the phone sequence of a vowel-to-vowel transition, the second key is the character pair that produces this transition. The value that is associated with these two keys is the occurence count of this character pair. So, if for instance the character pair *han kuk* occured 48,952 times in the text corpus, the hash table would have a key *("A N G U", "han kuk")* that would be associated with the value 48,952. For each first key in the hash table we also store the total sum of occurrences. This value is stored using "total" as second key. Thus, the total number of times that the pronunciation transition *A N G U* appears in the text material could be accessed as

[1] Today I go to Korea.

table.value(“A N G U”, “total”)[2].

For illustration purposes we list some example entries of the resulting hash table:

table.value(“A N G U”, “total”)	= 83,021
table.value(“A N G U”, “han kuk”)	= 48,952
table.value(“A N G U”, “san ku”)	= 314
table.value(“A N G U”, “pan kun”)	= 239
⋮	
table.value(“A N G U”, “than kuk”)	= 1
table.value(“iEO N Ph iEO”, “total”)	= 1,008
table.value(“iEO N Ph iEO”, “yeon phyeong”)	= 533
table.value(“iEO N Ph iEO”, “myeon phyeong”)	= 127
⋮	
table.value(“iEO N Ph iEO”, “myeon phyeo”)	= 1
⋮	

6.2.2 Unit Merging

The unit merging process is controlled by the following iterative procedure:

1. Choose the vowel-to-vowel transition t that has the highest occurrence count, i.e. the highest *total* value in the hash table:
   $t = \text{argmax}_{key}\ \{\text{table.value}(key,\ \text{“total”})\}$

2. Select a certain set P of one or more character pairs which produce this transition:
   $P \subseteq \{(p_{i1}, p_{i2}) \mid \text{table.isKey}(t, (p_{i1}, p_{i2}))\}$

3. Merge all character pairs in set P in the text corpus:
   $\forall\ (p_{i1}, p_{i2})\ \epsilon\ P : \text{mergeInCorpus}(p_{i1}, p_{i2})$

[2] We use pseudo code to describe operations on the table data structure: *table.value(a,b)* returns the value that is associated with the key pair (a, b), *table.isKey(a,b)* returns a boolean value indicating whether the key pair (a, b) is in the table, *table.delete(a,b)* removes the key (a, b) and the associated value from the table. Additionally, *mergeInCorpus(a,b)* names a method that merges all occurences of the syllable pair (a, b) in the text corpus.

4. Remove the respective entries from the hash table.
   $\forall\ (p_{i1}, p_{i2})\ \epsilon\ P : \text{table.delete}(t, (p_{i1}, p_{i2}))\}$

One can think of different stop criteria for this algorithm, e.g. perplexity or OOV rate based. We chose an OOV rate of 5% as the stop criterion. The JRTk can handle a maximum of 64k words in the recognition vocabulary. The OOV rate, however, is still below 1% for the resulting systems when the maximum vocabulary has been reached. Thus, in practice we use the vocabulary limit of 64k as stop criterion for the algorithm.

To preserve the eojeol boundaries we limit the unit merging process to unit pairs that lie in the same eojeol unit (*IntraEojeol*). But as the text data is not free of errors we consider a second variation where also units over eojeol boundaries may be merged (*InterEojeol*). This way, we also find out whether knowing the eojeol boundaries is an important source of information.

Two ways of selecting character pairs in step 2 of the above procedure are evaluated:

MergeAll Select *all* character pairs that produce the transition t:

$P = \{(p_{i1}, p_{i2}) \mid \text{table.isKeyPair}(t, (p_{i1}, p_{i2}))\}$

As the transition t between units disappears from the corpus, the acoustic confusability for this transition is eliminated.

MergeMax Select *the most frequent* character pair(s) that produce t:

$$\begin{aligned} P = \{(p_{i1}, p_{i2}) \mid\ & \text{table.isKeyPair}(t, (p_{i1}, p_{i2})) \wedge \\ & \text{table.value}(t, (p_{i1}, p_{i2})) \geq \\ & \text{table.value}(t, (p_{j1}, p_{j2})),\ \forall j \neq i\} \end{aligned}$$

With this approach, the transition t is not necessarily eliminated from the corpus. But, merging the most frequent character pair(s) reduces confusability and, in practice, many of the selected pairs are phrase-like pairs which to merge makes sense also from a language modeling point of view.

As it is not feasible to apply this algorithm to very large text corpora in RAM, the algorithm is split up into two passes: In the first pass we use the hash table to calculate a list of candidate pairs for merging. Then chunks of these candidates are merged in the large text corpus on disk. This process is repeated until a vocabulary of no more than 64k units is reached.

6.3 Speech Recognition Systems

Each of the *MergeAll* and the *MergeMax* approach are evaluated with both the *InterEojeol* and the *IntraEojeol* variation. This results in four systems. The naming convention for these systems is as follows: *MergeIntraAll* combines the *MergeAll* and *IntraEojeol* approaches. *MergeInterMax*, *MergeIntraAll* and *MergeIntraMax* are analogous.

The number of character pairs that were joined before the stop criterion was reached is 5,393 for *MergeIntraMax*, 19,951 for *MergeIntraAll*, 1,581 for *MergeInterMax* and 15,069 for *MergeInterAll*. See table 6.1. The *MergeMax* approach merges less unit pairs than the *MergeAll* approach. This is because the unit pairs merged with *MergeMax* have, on an average, a higher frequency in the text data than the *MergeAll* units and therefore the potential of creating new vocabulary units is higher for *MergeMax* pairs.

	Number of character pairs merged
MergeIntraMax	5,393
MergeIntraAll	19,951
MergeInterMax	1,581
MergeInterAll	15,069

Table 6.1: Number of character pairs that were merged.

Table 6.2 summarizes the corpus and vocabulary size of the four systems for the corpora *Train* and *PartChosun+Train*. The figures listed are in terms of units that result from the respective merging process. Interestingly, although *MergeMax* merges less different unit pairs than *MergeAll*, the *MergeMax* corpus size is smaller than that of the corresponding *MergeAll* system. Again, this is because the average frequency of each merged pair is higher in the *MergeMax* approach.

The OOV rates of the corpora *Test* and *Test-Utts* for the four systems are summarized in table 6.3. Both OOV rate and number of OOV words are calculated on both *Train* and *PartChosun+Train*. For all four systems the OOV rate using *PartChosun+Train* is around one percent. Such a low OOV rate is very good for large vocabulary speech recognition.

	PartChosun+ Train	Train
MergeIntraMax		
corpus size	4,032,575	174,387
vocabulary size	65,498	17,124
MergeIntraAll		
corpus size	4,707,499	203,623
vocabulary size	65,495	14,766
MergeInterMax		
corpus size	4,992,805	221,377
vocabulary size	65,479	10,924
MergeInterAll		
corpus size	5,466,419	239,158
vocabulary size	65,491	10,836

Table 6.2: Summary of merge based language model corpora.

The vocabulary growth on *PartChosun+Train* for all four merge based systems is shown in figure 6.1. The vocabulary grows significantly slower than on an eojeol based corpus. Also, the vocabulary grows faster for *MergeMax* than for *MergeAll*.

	Test	*Test-Utts*
MergeIntraMax		
Train	5.40% (1159)	6.08% (121)
PartChosun+Train	0.98% (210)	1.31% (26)
MergeIntraAll		
Train	4.61% (1147)	4.88% (111)
PartChosun+Train	1.14% (283)	1.14% (26)
MergeInterMax		
Train	2.77% (750)	2.78% (69)
PartChosun+Train	0.85% (231)	0.93% (23)
MergeInterAll		
Train	2.93% (850)	3.35% (89)
PartChosun+Train	0.96% (280)	1.13% (30)

Table 6.3: Summary of OOV rates and, in parantheses, OOV words.

The eight figures 6.2 to 6.5 and D.4 to D.7 show coverage characteristics of the four merge systems. They show for each system the self coverage of both *Train* and *PartChosun+Train* and their cross coverage on *Test* and *Test-Utts*. The coverage diagrams for the *Train* corpus are in the appendix. These diagrams show that the coverage for merge based units is more consistent than for eojeols; the self coverage and the cross coverage diagrams are

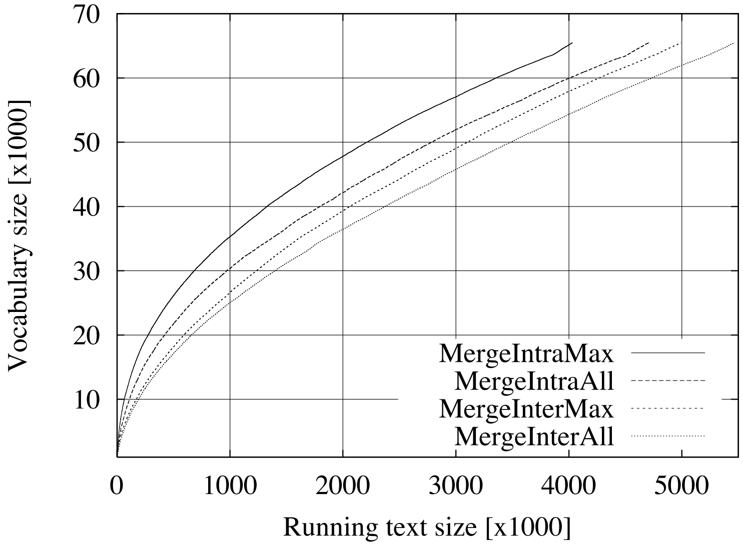

Figure 6.1: Vocabulary growth for the four merge based systems in corpus *PartChosun+Train*.

almost identical. For the eojeol units the cross coverage is up to five percent lower than the corresponding self coverage as can be seen from the diagrams 4.1 and 4.2. Comparing the figures shows us that coverage grows slightly slower in the *MergeAll* setup than in the *MergeMax* setup.

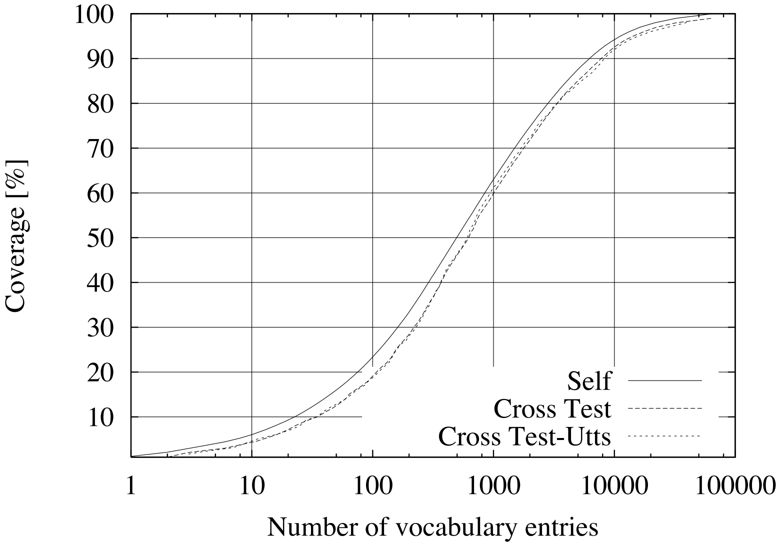

Figure 6.2: *MergeIntraMax*-based self coverage of *PartChosun+Train* and cross coverage of *Test* and *Test-Utts*.

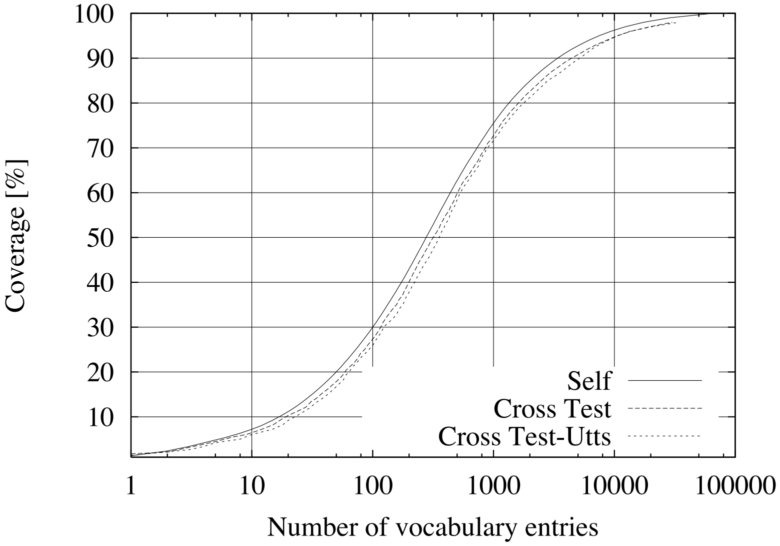

Figure 6.3: *MergeIntraAll*-based self coverage of *PartChosun+Train* and cross coverage of *Test* and *Test-Utts*.

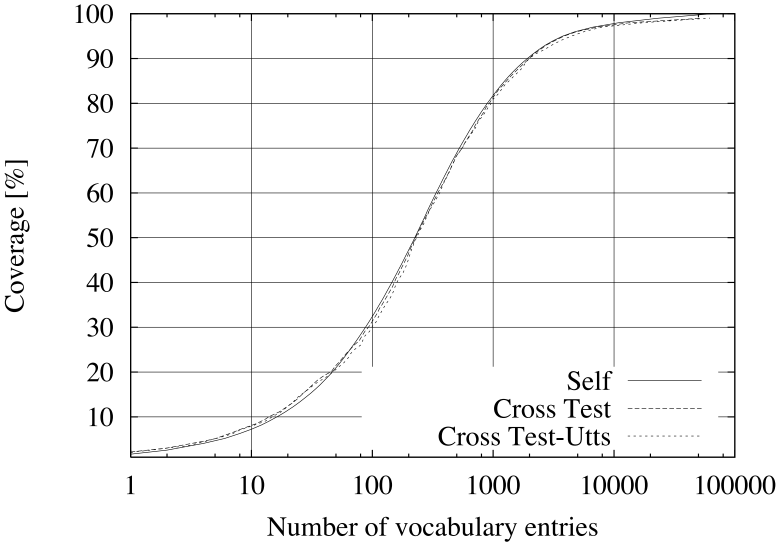

Figure 6.4: *MergeInterMax*-based self coverage of *PartChosun+Train* and cross coverage of *Test* and *Test-Utts*.

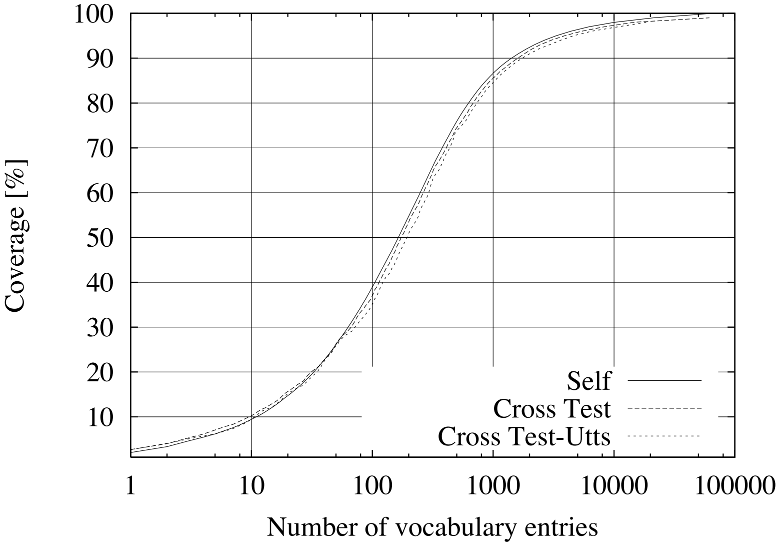

Figure 6.5: *MergeInterAll*-based self coverage of *PartChosun+Train* and cross coverage of *Test* and *Test-Utts*.

In addition to these four systems we also evaluated the *MergeIntraMax* approach, which performed the best of the four merge based systems, on the whole corpus *Chosun+Train*. This system will be referred to as *MergeIntraMaxWole*. With this approach 3,125 character pairs were merged. The characteristics of the resulting text corpora are as follows: *Chosun+Train* consists of 28,968,433 units and 65,328 vocabulary entries. The OOV rate and OOV words of *Test* are 0.12% and 28, respectively. For *Test-Utts* the according figures are 0.14% and 3. The *Train* corpus has a size of 196,185 units and a vocabulary size of 11,478. In this case the OOV rate and OOV words of *Test* are 2.84% and 682, respectively and for *Test-Utts* they are 3.01% and 67. Table 6.4 summarizes this information.

	Chosun+ Train	Train
Corpus characteristics		
corpus size	28,968,433	196,185
vocabulary size	65,328	11,478
OOV rates and OOV words		
Test	0.12% (28)	2.84% (682)
Test-Utts	0.14% (3)	3.01% (67)

Table 6.4: Summary of characteristics of *MergeIntraMax-Whole* corpora, and summary of OOV rates and, in parantheses, OOV words.

The vocabulary growth on *MergeIntraMaxWhole* is displayed in figure 6.6.

The figures 6.7 and D.8 show the coverage characteristics of the approach *MergeIntraMaxWhole* on the corpora *Chosun+Train* and *Train*.

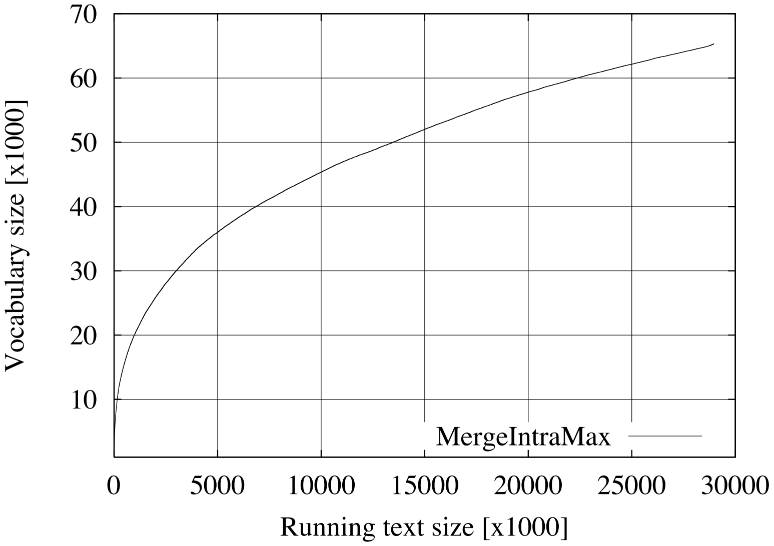

Figure 6.6: Vocabulary growth for *MergeIntraMaxWhole* based system in corpus *Chosun+Train*.

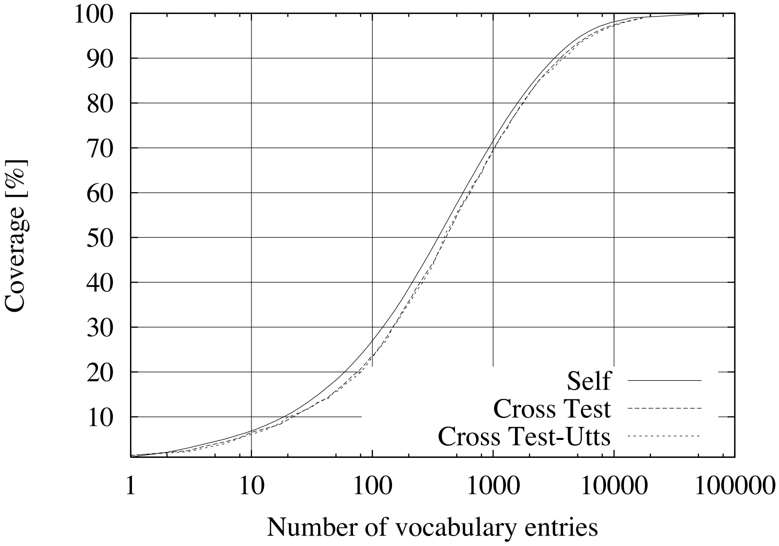

Figure 6.7: *MergeIntraMaxWhole*-based self coverage of *Chosun+Train* and cross coverage of *Test* and *Test-Utts*.

Chapter 7

Experiments

This chapter presents and discusses the results of the evaluated speech recognition systems. The necessary prerequisites for the experiments are discussed in section 7.1. Section 7.2 contains additional information on context dependent phone modeling. The experimental results are given in section 7.3.

7.1 Introduction

7.1.1 Recognition Accuracy

To determine the recognition performance of our systems, we measured the word error rate (WER), defined as:

$$\textit{Word Error Rate} = 100 \cdot \frac{\textit{substitutions + deletions + insertions}}{\textit{number of spoken words}}$$

As there is no straightforward Korean word unit error rates are given for several unit sets, namely eojeols, characters and phones. A value comparable to the word error rate in English is about the average of eojeol and character error rate. The following example illustrates the concept “error rate”. The upper line in the example shows the transcription of an utterance, the lower line shows the recognized hypothesis.

```
동무는 언제 아버님에게              편지를 씁니까
동무는      아버님과   어머님에게 편지를 씁니까
```

동무는 is recognized correctly. 언제 causes a deletion error as it does not appear in the hypothesis. 아버님에게 is confused with 아버님과. 어머님에게 is an insertion error in the hypothesis. The remaining two eojeols are recognized correctly. With five reference eojeols in total this leaves us an eojeol based error rate of $100 \cdot \frac{1+1+1}{5}\% = 60\%$.

To determine the character based error rate we break up each eojeol into its character components and calculate the error rate of the resulting two sequences of characters. In our example that is:

```
동 무 는 언 제 아 버 님             에 게 편 지 를 씁 니 까
동 무 는       아 버 님 과 어 머 님 에 게 편 지 를 씁 니 까
```

A total of 14 characters is recognized correctly. Two characters are deleted, four are inserted. Thus, the character error rate is $100 \cdot \frac{0+2+4}{16}\% = 37.5\%$.

To calculate the phone based error rate of these two sentences we look up the phonetic transcription of each eojeol unit in the dictionary and align the two resulting phone strings:

```
D O NG M U N EU N EO N J E A B EO N I M                        E G E ...
D O NG M U N EU N           A B EO N I M G O AE O M EO N I M E G E ...

...Ph iEO N J I R EU L SS EU M N I GG A

...Ph iEO N J I R EU L SS EU M N I GG A
```

The phone based error rate is $100 \cdot \frac{0+4+9}{36}\% = 36.11\%$.

7.1.2 Lattice Rescoring

Using Bayes' formula, acoustic probabilities and language model probabilities can be combined. Generally, the means and variances of these two classes of probabilities do not match and, as a consequence, either the acoustic or the language model side dominates the overall probability of a hypothesis. To compensate for this effect, the probabilities are weighted before they are multiplied. In addition, the JRTk decoder uses a word transition penalty q to control the length of the produced word sequence. The following variation of Bayes' formula is used:

$$P'(W \mid X) = \frac{P(X \mid W) \cdot P(W)^z \cdot q^{|W|}}{P(X)}$$

The exponent z is called the *language model weight*.

The use of z and q is very common in speech recognition research [44]. As it is not clear how to find the "optimal" (z, q) pair, the following strategy is used: A range of values is specified for both z and q. Then for each combination the overall recognition accuracy on the test set is calculated by rescoring the word lattice of each test utterance using these values. Finally, the pair (z_1, q_1) is picked which produces the highest recognition accuracy.

For our systems, we used the following heuristically determined ranges for z and q: $z \,\epsilon\, \{5, 10, \ldots, 55\}$ and $q \,\epsilon\, \{-30, -25, \ldots, 15\}$.

Our systems work with sub-eojeol units in the dictionary. However, we want to have a high-performance recognition on eojeol units. Thus, we must be able to determine which sub-eojeol units in a hypothesis have to be joined to form one eojeol unit. This issue is addressed as follows: each sub-eojeol unit that is inside an eojeol unit, as opposed to at the beginning, is prepended with a marker symbol, a dash. Then a sequence of sub-eojeol units can be transformed into a sequence of eojeols simply by connecting each "marked" unit to its predecessor. This applies to any type of sub-eojeol units we use; merge based units and morpheme units.

The lattice rescoring, i.e. determination of the "optimal" (z, q) pair, is then done as follows: for each combination of z and q we generate the eojeol versions of each sub-eojeol based reference and its respective hypothesis using the above strategy. On the resulting strings the eojeol word error rate can be calculated. We then pick the (z, q) pair that minimizes this eojeol error rate. For the speaker specific test results, minimal eojeol error rate is reported by determining a speaker-specific optimal (z, q) pair.

7.2 Context Modeling

Each phone is modeled context-dependently using a decision tree. These trees are generated using a set of 63 phonetically motivated context questions which are listed in table C.3 in the appendix.

In addition to the phonetically motivated questions we use two more questions that are related to the boundary characteristics of a phone. The first question asks whether a phone is at the boundary of a dictionary unit. This is a commonly used question in context-dependent modeling to capture cross-word effects. The second question asks whether a phone is at the left boundary of a dictionary unit which itself is not the left-most unit of an eojeol. This question gives the decoder component a mean to discriminate cross-unit models that are inside the same eojeol from cross-unit models over eojeol boundaries. In other words, this question helps the decoder to decide whether to stay within an eojeol unit or to start a new one.

7.3 Results And Discussion

7.3.1 Baseline

The complexity of a recognition task is generally measured in terms of its perplexity. The perplexity indicates the average word branching factor, i.e. the average number of words that can follow the current word. For an information theoretic discussion of the perplexity measure see [24]. Table 7.1 summarizes the perplexity information and also repeats the OOV rate for each system. These values are based on the corpora *PartChosun+Train* and *Test*. We see from the table that all merge based systems have significantly lower OOV rates than the morpheme based systems. The main reason is that the vocabulary size of the morpheme based corpora is much higher than 64k, see table 5.1. But only the most frequent 64k words can be used in the recognizer. On the other side, the merge based corpora contain no more than 64k different units. Unfortunately, the perplexity values displayed in table 7.1 are not directly comparable as they are calculated on different unit sets.

Table 7.2 shows the error rates for the four merge based systems and the two

	OOV	Perplexity
MergeIntraMax	0.943%	481.7
MergeIntraAll	1.102%	206.3
MergeInterMax	0.830%	149.5
MergeInterAll	0.939%	104.7
Morph	2.371%	209.2
MorphTag	2.784%	216.4

Table 7.1: Recognition task complexities on *Chosun+Train*.

morpheme based systems. Lattice rescoring is done on the eojeol level. Character and phone error rate is then determined as described in section 7.1.1 with the same (z, q) pair. Note, that the reported character and phone error rates do not necessarily have to be maximal using this evaluation scheme. In the first column, the table shows the error rates on the whole test set. The columns labelled "1" to "10" show the speaker-specific error rates. These results were obtained using the corpus data *PartChosun+Train*.

We see from table 7.2 that the plain morpheme system *Morph* is not outperformed by *MorphTag*. The appended POS-tag gives the *MorphTag* units additional information, but this also increases the size of the vocabulary which leads, due to the 64k test vocabulary limitation, to a higher OOV rate than for the system *Morph*. Also, the language model is less reliable as more 3-gram models have to be estimated from the same amount of text data. In combination, these effects make both morpheme systems perform equally well.

The performance of *MergeInterMax* and *MergeInterAll* is significantly worse than that of either the *IntraEojeol* based systems or the morpheme systems. This tells us that the eojeol boundaries are an important structural element and that merging units over eojeol boundaries seems to destroy this structure, leaving us with units which are not well suited for 3-gram modeling.

We see from table 7.2 that *MergeMax* outperforms *MergeAll* in both the *InterEojeol* and the *IntraEojeol* case. There are two reasons for this behavior: First, the character pair which is merged is in many cases part of a suffix combination or a multiple-character stem that occurs very frequently. These

	All	1	2	3	4	5	6	7	8	9	10
MergeIntraMax											
Phone ER	16.2	19.3	21.3	12.3	10.3	13.4	16.0	12.6	18.1	15.5	27.4
Character ER	24.2	27.4	28.6	17.0	17.0	19.4	23.2	20.1	26.8	22.8	37.1
Eojeol ER	39.8	41.4	43.2	24.4	31.9	35.0	37.3	32.5	38.6	38.8	64.8
MergeIntraAll											
Phone ER	17.9	20.9	21.3	12.5	11.1	14.4	15.5	13.9	17.8	17.0	27.2
Character ER	25.9	30.4	28.4	18.8	17.9	21.4	22.8	21.6	26.8	24.4	38.6
Eojeol ER	44.2	48.1	48.6	31.5	34.8	37.5	39.1	36.8	38.6	42.9	59.2
MergeInterMax											
Phone ER	19.9	21.7	22.8	14.7	13.0	14.3	20.1	18.1	20.5	17.4	23.2
Character ER	29.5	32.7	35.3	21.7	20.9	19.9	30.5	28.4	32.1	25.2	32.2
Eojeol ER	48.6	52.6	60.6	33.1	37.7	41.7	46.4	45.6	49.1	42.1	57.7
MergeInterAll											
Phone ER	20.2	21.3	22.9	12.2	16.6	11.4	16.1	20.4	27.6	19.7	24.6
Character ER	29.7	31.0	32.4	17.3	25.4	17.4	24.4	33.3	38.0	29.4	37.6
Eojeol ER	50.0	50.0	60.2	32.3	42.8	33.3	40.9	52.6	57.9	50.0	57.7
Morph											
Phone ER	15.6	20.0	17.4	10.8	11.6	10.6	13.9	13.1	20.2	12.9	20.8
Character ER	22.4	29.5	24.9	15.5	17.9	14.5	21.5	18.2	26.8	18.9	30.7
Eojeol ER	38.1	42.9	43.1	26.0	30.4	28.3	35.5	32.5	44.7	31.7	52.1
MorphTag											
Phone ER	15.8	17.3	19.0	13.9	12.6	10.3	14.0	14.5	17.5	14.4	24.8
Character ER	23.3	27.1	27.1	19.6	20.9	15.1	21.2	22.5	24.0	20.7	34.2
Eojeol ER	38.2	39.1	41.4	30.7	34.1	29.2	33.6	38.6	41.2	33.3	59.2

Table 7.2: Summary of recognition error rates, %.

are units that would be combined in a phrase based language model and merging them thus makes sense from a language modeling perspective. The *MergeMax* approach only merges such max-pairs, therefore generating more phrase-like units. This seems to increase the language model suitability of the resulting units. Second, merging every pair that produces a pronunciation transition, as done in the *MergeAll* approach, eliminates this transition totally. But at the same time this increases the size of the vocabulary, decreasing the number of further transitions to merge.

The most interesting result that we can extract from table 7.2 is that the system *MergeIntraMax* and the morpheme based system perform about equally well. The speaker-specific error rates show that on some speakers the morpheme systems perform better while on others they are worse than *MergeIntraMax*. In fact, the total eojeol error rate of the morpheme system is about 1.5% smaller than for the merge based system. But as the eojeol error rate is a very sensitive measure[1], we can not imply that the morpheme systems

[1] Consider for example the eojeol sequence "a-b-c d-e f-g", where the letters symbolize

perform better than *MergeIntraMax* in general.

7.3.2 Pronunciation Variant LM

We can further improve the recognition performance by using a "pronunciation-variant based language model". As said in section 4.3, the pronunciation of a unit depends on its context. Different pronunciations of a unit are handled as pronunciation variants in the dictionary, distinguished by appended variant indizes. To estimate a variant based language model, each unit in the text corpus gets appended the index of its pronunciation variant in the dictionary. The 3-gram model is calculated on the resulting text corpus. Thus, pronunciation variants of a unit are considered different language model units. One disadvantage is that the number of units is increased and therefore more 3-grams have to be estimated from the same amount of text material. In addition, these variant language models have slightly higher perplexities and OOV rates than their counterparts that do not distinguish pronunciation variants. Each pronunciation variant of a unit is considered a distinctive vocabulary unit in the recognizer. With the 64k vocabulary size limitation less different words can be used in the vocabulary and as a consequence the OOV rate will increase. We evaluated this idea on the systems *MergeIntraMax* and *MorphTag*. The task perplexity and the OOV rate of these systems are given in table 7.3.

	OOV	Perplexity
MergeIntraMax	1.518%	498.3
MorphTag	3.575%	224.4

Table 7.3: Recognition task complexities on *Chosun+Train*, pronunciation variants not mapped on baseform in language model.

Table 7.4 shows the performance results. Introducing a variant-LM decreases the eojeol error rate of the system *MergeIntraMax* to 35.6%, an error reduc-

syllable components. While both the character and phone error rate for "a-b c-d-e f-g" are 0%, the eojeol error rate would be 66%. Small errors can yield big differences in the eojeol error rate.

tion of 10.5%. For *MorphTag* the eojeol error rate is reduced by 15.4% to 32.3%.

	All	1	2	3	4	5	6	7	8	9	10
MergeIntraMax											
Phone ER	14.3	15.2	16.4	8.9	10.7	10.1	13.4	9.6	18.3	13.0	22.7
Character ER	21.2	24.2	24.3	11.4	17.0	15.7	20.3	15.1	24.9	17.6	33.2
Eojeol ER	35.6	36.1	41.5	21.3	30.4	25.8	34.5	27.2	36.0	29.5	59.2
MorphTag											
Phone ER	13.5	14.2	18.5	10.0	9.7	7.5	12.0	8.2	12.0	11.1	22.2
Character ER	19.8	21.2	26.9	14.7	14.5	10.8	16.4	12.7	16.8	16.0	33.7
Eojeol ER	32.3	33.1	40.6	25.2	24.6	21.7	26.4	22.8	28.1	30.9	57.7

Table 7.4: Summary of recognition error rates using a variant-LM, %.

7.3.3 More Data

In addition, we evaluated the two above systems, using the variant-LM, on the full text corpus *Chosun+Train*. The resulting systems are called *MergeIntraMaxWhole* and *MorphTagWhole*. Using a larger amount of text data improves the robustness of the language model. To keep the vocabulary size of the merge system at 64k, the merging has to be done on the bigger text corpus again. The recognition task characteristics of the resulting systems are shown in table 7.5. The OOV rate and perplexity values show us that the merge based system profits a lot from the enlarged amount of text data. The OOV rate is reduced by more than one percent and the perplexity is almost reduced by a factor of five. But again, the perplexity values cannot be compared directly as the unit sets of the two *MergeIntraMax* systems differ. This is because less units can be merged on the big corpus before a 64k vocabulary is reached. The OOV rate of *MorphTag* remains around 3.5% but the perplexity increases by a factor of two. The reason for this is the increase in vocabulary size by a factor of about four.

The performance results of *MergeIntraMaxWhole* and *MorphTagWhole* are listed in table 7.6. The increased amount of text data results in an eojeol error rate of 28.0% for *MergeIntraMaxWhole*, an error reduction of 21.3%. For *MorphTagWhole* the eojeol error rate is reduced by 3.1% to 31.3%. While

	OOV	Perplexity
MergeIntraMaxWhole	0.233%	136.6
MorphWhole	2.886%	143.1
MorphTagWhole	3.513%	485.9

Table 7.5: Recognition task characteristics.

the merge based system profits from the increased amount of text data, the improvement for the morpheme system is not very significant. The main reason for this difference in improvement should be that the vocabulary for *MergeIntraMaxWhole* is kept at 64k resulting in a very low OOV rate of 0.23%.

	All	1	2	3	4	5	6	7	8	9	10
MergeIntraMaxWhole											
Phone ER	11.6	14.7	14.7	7.3	9.1	9.2	12.4	10.3	12.3	9.1	19.9
Character ER	16.4	20.4	21.0	9.7	14.8	12.5	17.4	15.1	17.1	12.3	29.7
Eojeol ER	28.0	29.3	33.7	16.4	26.1	22.5	27.3	25.4	24.6	22.3	47.9
MorphTagWhole											
Phone ER	12.2	10.6	16.8	9.4	11.2	6.4	12.9	10.8	13.5	9.1	18.5
Character ER	18.1	14.7	24.5	13.5	18.2	10.0	18.6	18.5	18.4	12.9	26.2
Eojeol ER	31.3	23.1	38.9	25.0	32.1	19.8	31.8	29.8	29.8	24.5	47.9

Table 7.6: Summary of recognition error rates on *Chosun+Train*, %.

7.3.4 Corrected Speech Database

An analysis of the recognition errors revealed that one major type of error was made, namely the deletion of one syllable in the beginning or the end of an utterance. As a consequence, the segmentation of the audio waveform files was checked and it was found that most of them were segmented too sharply. In many cases, as much as a whole syllable was cut away. Subsequently, the whole database was resegmented. New labels were created with the repaired database and the acoustic models were retrained on these new labels. With the new acoustic models we repeated the last experiment on the new database. In addition to *MergeIntraMaxWhole* and *MorphTagWhole* the system *MorphWhole* was also evaluated.

Table 7.7 shows the performance results. Working on the corrected database decreases the eojeol error rate of *MergeIntraMaxWhole* to 24.6%, an error reduction of 12.1%. For *MorphTag* the eojeol error rate is reduced by 4.1% to 30.0%. *MorphWhole* has an eojeol error rate of 24.0%

	All	1	2	3	4	5	6	7	8	9	10
MergeIntraMaxWhole											
Phone ER	9.9	11.4	11.3	8.0	6.7	7.5	11.8	8.2	9.6	6.6	12.9
Character ER	14.5	18.6	15.5	11.4	10.1	12.0	16.4	11.1	12.8	10.2	20.3
Eojeol ER	24.6	24.8	29.1	16.5	20.3	21.7	26.4	21.1	18.4	18.6	29.6
MorphWhole											
Phone ER	9.4	5.8	13.0	4.8	6.5	9.0	11.7	8.9	8.9	7.0	15.9
Character ER	13.0	7.4	18.2	6.7	10.3	11.4	18.6	11.1	10.6	10.0	22.3
Eojeol ER	24.0	13.4	34.3	14.1	16.7	20.8	28.2	18.4	17.5	18.0	38.0
MorphTagWhole											
Phone ER	10.7	6.7	14.2	8.2	9.7	8.5	11.5	7.5	9.4	8.2	17.1
Character ER	16.2	11.2	21.8	12.3	14.0	13.1	18.0	13.0	12.5	12.1	21.8
Eojeol ER	30.0	19.4	37.1	21.9	27.0	27.5	35.5	27.2	21.9	21.1	42.3

Table 7.7: Summary of recognition error rates on *Chosun+Train* with corrected speech database, %.

7.3.5 Discussion

The best merge based baseline system, *MergeIntraMax*, has an eojeol error rate of 39.8%. The error rate of the system *MorphTag* is 1.6% lower, 38.2%. Introducing pronunciation-variant based language models reduces the eojeol error rate of *MergeIntraMax* by 10.5% to 35.6%. The error rate of *MorphTag* using a variant based language model is 32.3%, an error reduction of 15.4%. Increasing the amount of training data for the language model yields a further error reduction of 21.3% for *MergeIntraMax*, but only 3.1% for *MorphTag*. The resulting eojeol eror rates are 28.0% for *MergeIntraMax* and 31.3% for *MorphTag*. We found out that most utterances in the speech database were segmented too sharply. As a consequence, we corrected the segment boundaries of each utterance and used the repaired database to retrain the acoustic models. Using the variant-LM and the large text corpus on the new database, *MergeIntraMax* has an eojeol error rate of 24.6%, *MorphTag* has an error rate of 30.0%. This is an error reduction of 12.1% and

4.1%, respectively. To summarize, we reduced the eojeol error rate of 39.8% of the baseline *MergeIntraMax* system to 24.6%, a total error reduction of 38.2%. The eojeol error rate of the baseline *MorphTag* system, which was 38.2%, was reduced by 21.5% to an error rate of 30.0%. This information is summarized in table 7.8

The system *Morph* has an eojeol error rate of 24.0% on the corrected database. It outperforms the system *MorphTag* due to two reasons. First, the vocabulary size of *MorphTag* is significantly higher than the one of *Morph.* As a consequence, *MorphTag* has a higher OOV rate when its vocabulary is limited to 64k. Also, the *MorphTag* language model is less robust than the *Morph* language model as more 3-grams have to be estimated from the same amount of training data.

	MergeIntraMax		*MorphTag*	
	Eojeol Error Rate	Error Reduction	Eojeol Error Rate	Error Reduction
Baseline	39.8	-	38.2	-
Variant-LM	35.6	10.5	32.3	15.4
More Data	28.0	21.3	31.3	3.1
Corrected Database	24.6	12.1	30.0	4.1

Table 7.8: Summary of the performance of the recognition systems.

All these experiments consistently show that the dictionary units determined by the data-driven approach *MergeIntraMax* are no worse than the morpheme units determined by a complex morpheme analyzing system.

Chapter 8

Conclusions

8.1 Conclusions

The “agglutinating” nature of the Korean language makes the choice of appropriate dictionary units for a large vocabulary speech recognition system difficult. Eojeols, the units that result from the agglutination process, are highly inappropriate. They have an inherent severe OOV problem and show a linear with task size vocabulary growth rate. Sub-eojeol units must therefore be determined. A natural choice of sub-eojeol unit is the morpheme as each eojeol is composed of a sequence of morphemic components. The development of a system that automatically splits eojeol units into their morpheme components is very expensive, involving a lot of *apriori* expert knowledge on the morphological structure of the Korean language.

This work describes a new approach to finding appropriate dictionary units. Because morpheme units are not proven to be the optimal choice of sub-eojeol units from a language modeling perspective we discarded the morphemic structure and approached the problem from an acoustic, data-driven perspective instead. We developed an approach which, in a first pass, splits each eojeol into its syllable components. Then, iteratively, syllable pairs are merged in order to reduce acoustic confusability of syllable transitions.

Our results show that this data-driven approach generates units that are equally suitable for a large vocabulary speech recognition dictionary as the morpheme units. As a consequence, we do not have to rely on complex expert systems for the unit determination process. It can be done automatically using a data-driven approach.

8.2 Future Work

To further improve the performance of these speech recognition systems, a number of possibilities might be explored:

- *"LM-related" merging criterion:*
  A language model related extension of the function that chooses the unit pairs to merge could improve the quality of the determined unit set in terms of language model suitability. A simple example of such an extension would be the integration of *mutual information.*
  It would also be interesting to compare the unit merging approach to a purely LM motivated approach where iteratively the most frequent pair of syllables is merged, without taking into account the frequency of phone transitions.

- *"More Data":*
  A famous expression in the area of pattern recognition is: "There is no data like more data". One way to improve the recognition performance would be to collect more speech data for the acoustic models. Not only can the models be estimated more reliably but the system also increases its speaker independent robustness by learning more about "new speakers". Analogous, more text data would improve the quality of the statistical language model.

- *Higher-order LMs:*
  The employment of higher order n-gram language models ($n > 3$) would very likely yield an increase in performance. As the language model is based on sub-eojeol units, it is not always guaranteed that the 3-gram provides enough range to ensure valid eojeol unit connection.

- *System Tuning:*
  A general way to improve the performance of a speech recognition system is to tune its parameters. This would include the number of polyphone models, the dimension of the feature vector, the number of Gaussians and also variations in the preprocessing.

Appendix A

Transcription Systems for 한글

한글	McCune-	Yale	North	South	hcode
	Reischauer		Korea	Korea	
ㅂ	p,b	p	p	b	p
ㅍ	p’	ph	ph	p	ph
ㅃ	pp	pp	pp	bb	pp
ㄷ	t,d	t	t	d	t
ㅌ	t’	th	th	t	th
ㄸ	tt	tt	tt	dd	tt
ㅅ	s	s	s	s	s
ㅆ	ss	ss	ss	ss	ss
ㅈ	ch,j	c	ts	j	c
ㅊ	ch’	ch	tsh	ch	ch
ㅉ	tch	cc	tss	jj	cc
ㄱ	k,g	k	k	g	k
ㅋ	k’	kh	kh	k	kh
ㄲ	kk	kk	kk	gg	kk
ㅁ	m	m	m	m	m
ㄴ	n	n	n	n	n
ㅇ	ng	ng	ng	ng	ng
ㅎ	h	h	h	h	h
ㄹ	l,r	l	r	l,r	l,r
ㅣ	i	i	i	i	i

Table A.1: Transcription systems for the 한글 consonants.

한글	McCune-Reischauer	Yale	North Korea	South Korea	hcode
ㅣ	i	i	i	i	i
ㅟ	wi	wi	wi	wi	wi
ㅔ	e	ey	e	e	e
ㅖ	ye	yey	ye	ye	ye
ㅞ	we	wey	we	we	we
ㅚ	oe	oy	oi	oe	oe
ㅐ	ae	ay	ai	ae	ae
ㅒ	yae	yay	yai	yae	yae
ㅙ	wae	way	wai	wae	wae
ㅡ	ŭ	u	ŭ	eu	eu
ㅓ	ŏ	e	ŏ	eo	eo
ㅕ	yŏ	ye	yŏ	yeo	yeo
ㅝ	wŏ	we	wŏ	weo	weo
ㅏ	a	a	a	a	a
ㅑ	ya	ya	ya	ya	ya
ㅘ	wa	wa	wa	wa	wa
ㅜ	u	wu	u	u	u
ㅠ	yu	yu	yu	yu	yu
ㅗ	o	o	o	o	o
ㅛ	yo	yo	yo	yo	yo
ㅢ	ui	uy	ui	eui	yi

Table A.2: Transcription systems for the 한글 vowels.

Appendix B

Text Corpus Mappings

Acronym	한글	Acronym	한글
3D	쓰리디	IBM	아이비엠
ABC	에이비씨	ISDN	아이에스디엔
ADAC	에이디에이씨	JPEG	제이펙
AI	에이아이	LAPD	엘에이피디
AMD	에이엠디	MBA	엠비에이
ATM	에이티엠	MRI	엠알아이
AT&T	에이티앤드티	NYSE	엔와이에스이
AWACS	어웩스	OEM	오이엠
BASF	비에이에스에프	PCMCIA	피씨엠씨아이에이
CDU/CSU	씨디유 씨에스유	PDA	피디에이
CeBIT	씨비트	R&D	알앤드디
CEO	씨이오	RAM	램
CGI	씨지아이	TOEFL	토플
CIA	씨아이에이	UCLA	유씨엘에이
CMOS	씨모스	UEFA	유이파
CPU	씨피유	UN	유엔
DAEWOO	대우	UNESCO	유네스코
EPROM	이프롬	UNICEF	유니쎄프
FBI	에프비아이	US	유에스
FIFA	피파	WTO	더블유티오
HBO	에이치비오	YMCA	와이엠씨에이

Table B.1: Some examples of acronym mappings.

Descriptor	한글	Descriptor	한글
%	퍼센트	m^3	세제곱미터
cc	씨씨	m^2	제킬로미터
cm	쎈티미터	m	미터
dB	데시벨	MB	메가바이트
g	그람	MHz	메가헤르쯔
GB	기가바이트	mm	밀리미터
ha	헥타	mW	밀리와트
Hz	헤르쯔	kW	킬로와트
kbps	킬로비피에스	pH	피에이치
kg	킬로그람	ppm	피피엠
kHz	킬로헤르쯔	t	톤
km	킬로미터	V	볼트
kV	킬로볼트		

Table B.2: Mapping table for units.

Appendix C

Janus Toolkit

C.1 Phoneme Models

IPA- symbol	Janus- Phone	Example
a	A	sh**ah**
ɛ	AE	**ca**t
e	E	b**e**t
i	I	**i**nn
o	O	**o**il
ə	EO	**a**bout
ø	OE	German **Goe**the, but also **we**t
u	U	g**oo**se, but shorter (rounded lips)
ɨ	EU	brok**e**n (unrounded lips)
y	UE	German f**ü**nf
ɨi	euI	sq**uee**ze pedantically as [ɨi] but usually initial as [ɨ], in other position as [i], and as particle meaning “of” as [e]
ia	iA	**ya**rn
ie	iE	**ye**llow
iə	iEO	**ye**rba
io	iO	**yo**ke
iu	iU	**you**th
oa	oA	**wa**sh
uə	uEO	**wo**nderful

Table C.1: Korean vowel models.

IPA- Symbol	Janus- Phone	Example and explanation
č	CHh	hit**chh**ike, strong aspiration
ǰ	J	re**j**oice, fully voiced
č’	JJ	pi**tch**er, tight throated release
s	S	**s**ing, weaker than its English counterpart
s’	SS	a**ss**ail, with tension in the throat and tongue
m	M	**m**omentum
n	N	**n**ewtonian, tongue tip behind upper teeth
ŋ	NG	si**ng**ing
h	H	**h**at
p	Ph	u**ph**ill, strong aspiration
b	B	o**b**ey, fully voiced
p˺	p	sto**p**, unreleased
p’	BB	s**p**in, tight throated release
t	Th	ho**th**ouse, strong aspiration
d	D	a**d**o, fully voiced
t˺	t	ye**t**, unreleased
t’	DD	s**t**ay, tight throated release
k	Kh	blo**ckh**ead, strong aspiration
g	G	a**g**o, fully voiced
k˺	k	ti**c**, unreleased
k’	GG	s**k**y, tight throated release
ɾ	R	ve**r**y (in British English)
l	L	tai**l**

Table C.2: Korean consonant models.

C.2 Phoneme Context Question Sets

CONSONANTS	Ph B BB p Th D DD t Kh G GG k CHh J JJ M N NG R L S SS H		
VOWEL	A EO O U I EU AE E OE UE iA iEO iO iU iE oA uEO euI		
VCD	B BB D DD G GG J JJ NG M N R L		
VELAR	Kh G GG k NG	SILENCE	SIL
BILABIAL	Ph B BB p	NOISE	+hGH
PALATAL	CHh J JJ	NASAL	M N NG
AFFRICATE	CHh J JJ	LATERALAPPR	L
EMPHASED	BB DD SS JJ GG	FRICATIVE	S SS H
ASPIRED	Ph Th Kh CHh	FRIC-ALVEO	S
PLOS-VCD	B BB D DD G GG	FRIC-GLOTTAL	H
PLOS-UNVCD	Ph Th Kh	EMPH-UNVCD	SS
CLOSE-VOW	I UE EU U	EMPH-BILABIAL	BB
EMPH-VCD	BB DD GG JJ	EMPH-VELAR	GG
VOW-I	I iA iEO iO iU iE	EMPH-ALVEO	DD SS
ALVEO-VCD	D DD N R L	EMPH-PALATAL	JJ
ALVEO-UNVCD	Th t S SS	IMPL-BILABIAL	p
DIPH-I	iA iEO iO iU iE	IMPL-ALVEO	t
ROUND	O U OE UE	IMPL-VELAR	k
UNROUND	EO EU I A AE E	BILABIAL-VCD	B BB M
FRONT-VOW	I E AE A UE OE	BILABIAL-UNVCD	Ph p
BACK-VOW	EU EO O U	PALATAL-VCD	J JJ
VOW-U	U UE uEO	PALATAL-UNVCD	CHh
ALVEO	Th D DD t N R L S SS	VELAR-UNVCD	Kh k
PLOSIVE	Ph B BB Th D DD Kh G GG	DIPH-O	oA
TRILL	R	DIPH-U	uEO
UNVCD	Ph p Th t Kh k CHh S SS H	DIPH-EU	euI
DIPHTHONG	iA iEO iO iU iE oA uEO euI	DIPH-DOWN	euI
DIPH-UP	iA iEO iO iU iE oA uEO	OPENMID-VOW	AE EO
VOW-O	O OE oA	OPEN-VOW	A
PLOS-BILABIAL	Ph B BB	VOW-EU	EU euI
PLOS-ALVEO	Th D DD	AFFR-VCD	J JJ
PLOS-VELAR	Kh G GG	AFFR-UNVCD	CHh
VELAR-VCD	G GG NG	IMPLOSIVE	p t k
GLOTTAL	H	CLOSEMID-VOW	E OE O

Table C.3: Phone sets used for decision tree based context dependent phone modeling.

Appendix D

Coverage

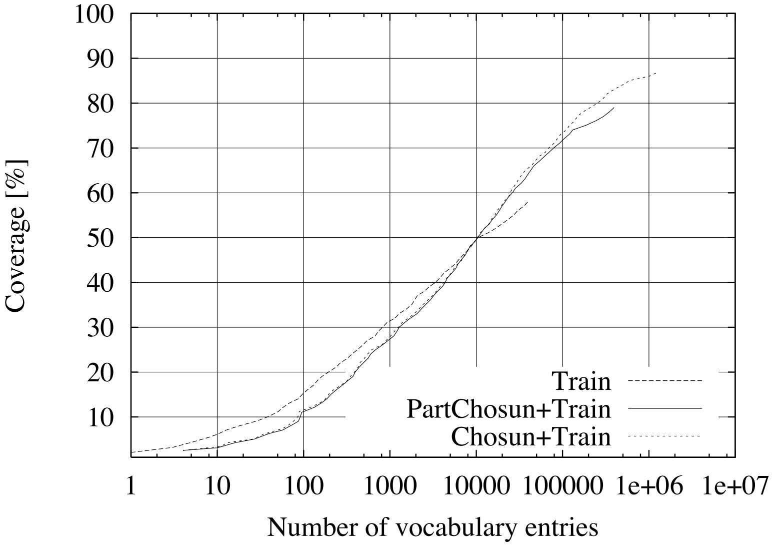

Figure D.1: Eojeol-based cross coverage of *Test-Utts* with different language model corpora.

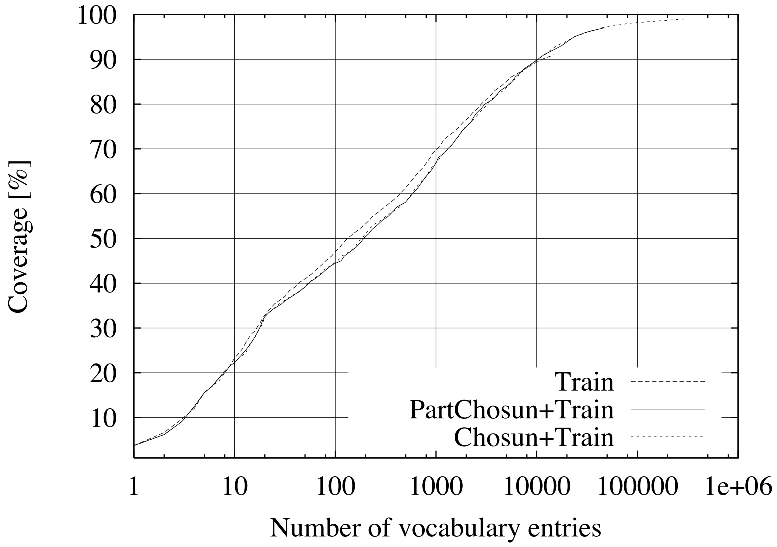

Figure D.2: *Morph*-based cross coverage of *Test-Utts* with different language model corpora.

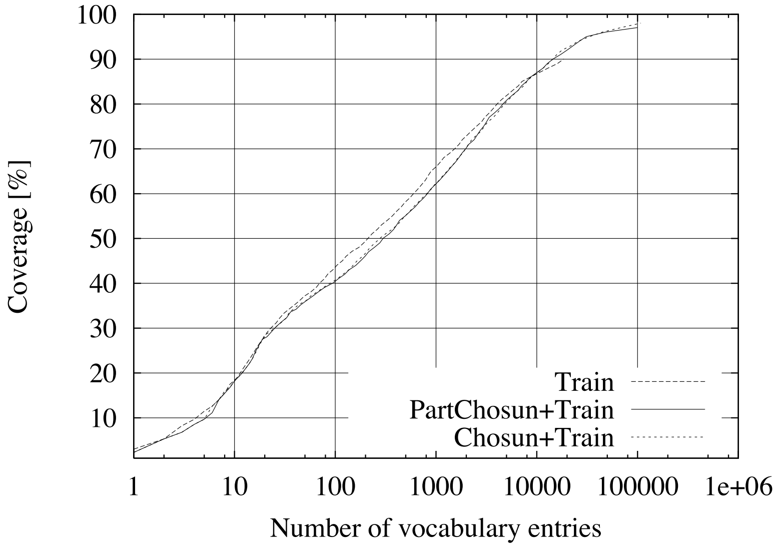

Figure D.3: *MorphTag*-based cross coverage of *Test-Utts* with different language model corpora.

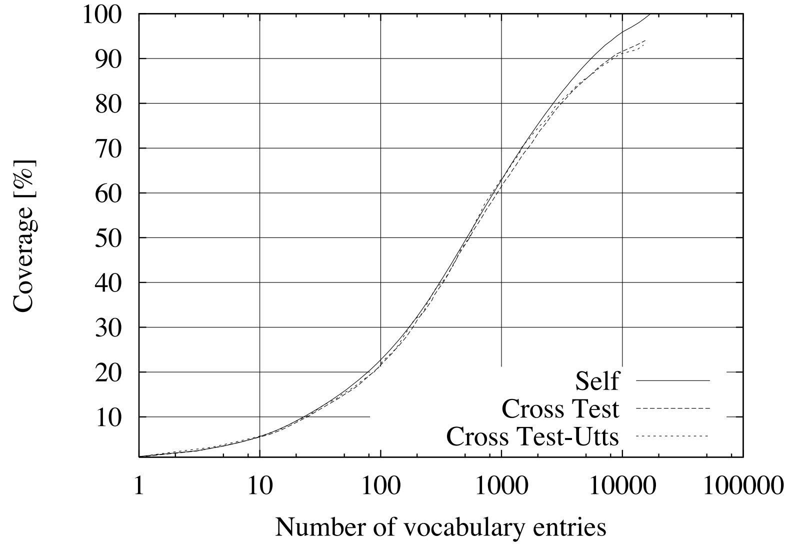

Figure D.4: *MergeIntraMax*-based self coverage of *Train* and cross coverage of *Test* and *Test-Utts*.

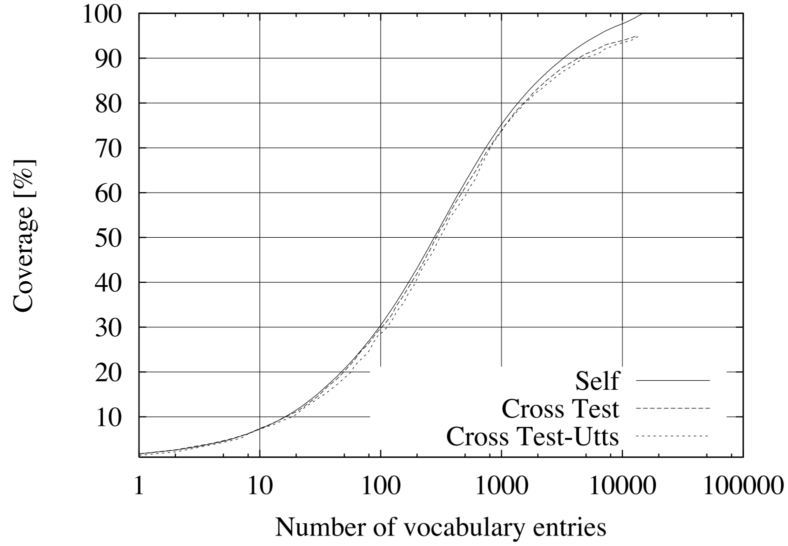

Figure D.5: *MergeIntraAll*-based self coverage of *Train* and cross coverage of *Test* and *Test-Utts*.

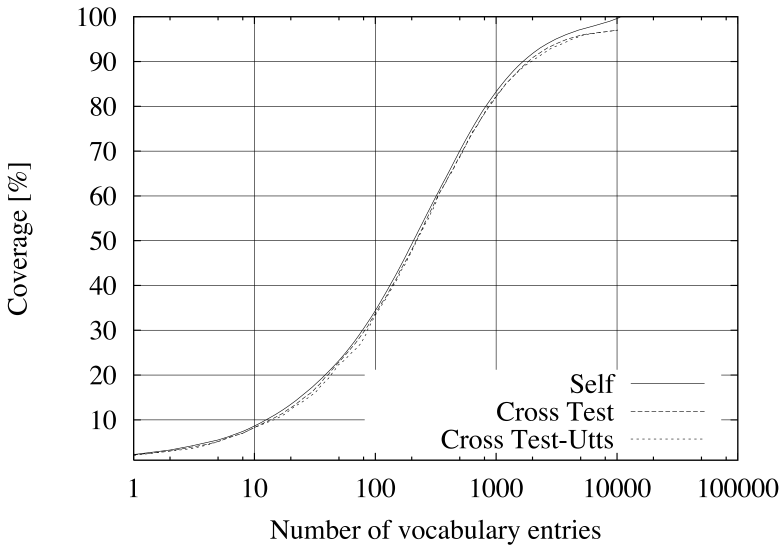

Figure D.6: *MergeInterMax*-based self coverage of *Train* and cross coverage of *Test* and *Test-Utts*.

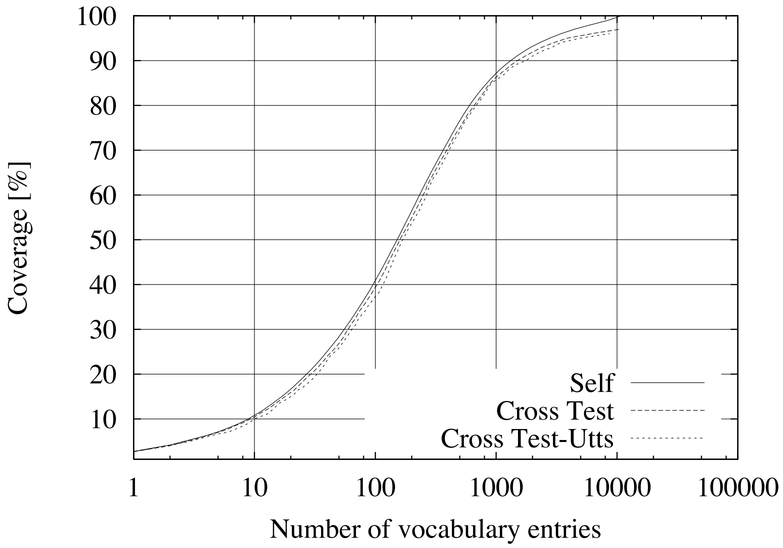

Figure D.7: *MergeInterAll*-based self coverage of *Train* and cross coverage of *Test* and *Test-Utts*.

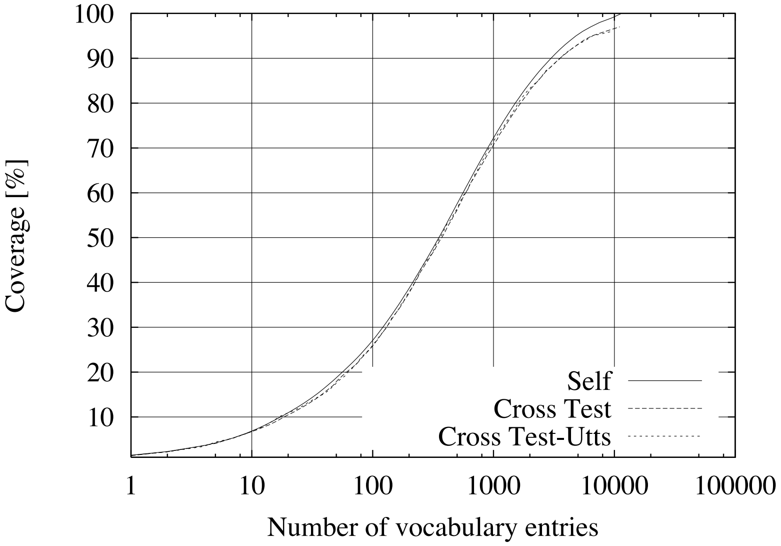

Figure D.8: *MergeIntraMaxWhole*-based self coverage of *Train* and cross coverage of *Test* and *Test-Utts*.

Bibliography

[1] International Phonetic Association. *Handbook of the International Phonetic Association*. Cambridge University Press, 1999.

[2] X. Aubert, R. Haeb-Umbach, and H. Ney. Continuous Mixture Densities and Linear Discriminant Analysis for Improved Context-Dependent Acoustic Models. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 648–651. IEEE Signal Processing Society, IEEE; New York, NY, April 1993.

[3] L.R. Bahl, R. Bakis, P.S. Cohen, A.G. Cole, F. Jelinek, B.L. Lewis, and R.L. Mercer. Further Results on the Recognition of a Continuously Read Natural Corpus. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 872–876. IEEE, IEEE; New York, NY, April 1980.

[4] L.E. Baum. An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes. *Inequalities*, (3):1–8, 1972.

[5] G. Boulianne and P. Kenny. Optimal Tying of HMM Mixture Densities Using Decision Trees. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP 96*, volume 1, pages 350–353. IEEE Signal Processing Society, IEEE; New York, NY, October 1996.

[6] E.O. Brigham. *FFT : schnelle Fourier-Transformation*. R. Oldenburg Verlag, 1995.

[7] J.C. Buckow. Klassifizierung und Erkennung von Sprachsegmenten. Master's thesis, Universität Karlsruhe, October 1996.

[8] P.A. Chou. Optimal Partitioning for Classifications and Regression Trees. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 13(4):340–354, April 1991.

[9] K.Y.O. Dowling. *Korean Phrasebook.* Lonely Planet, 1995.

[10] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, 1973.

[11] E. Eide and H. Gish. A Parametric Approach to Vocal Tract Length Normalization. In *Proceedings of the IEEE International Conference On Acoustics, Speech And Signal Processing*, volume 1, pages 346–348. IEEE Signal Processing Society, IEEE; New York, NY, May 1996.

[12] M. Finke, J. Fritsch, P. Geutner, K. Ries, and A. Waibel. The JanusRTk Switchboard/Callhome 1997 Evaluation System. In *Proceedings of the LVCSR Hub5-e Workshop*, May 1997.

[13] M. Finke and I. Rogina. Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1743–1746. IEEE Signal Processing Society, IEEE Comput. Society Press; Los Alamitos, CA, April 1997.

[14] J. Fritsch and M. Finke. ACID/HNN: Clustering Hierarchies of Neural Networks for Context-Dependent Connectionist Acoustic Modeling. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 505–508. IEEE Signal Processing Society, IEEE; New York, NY, May 1998.

[15] J. Fritsch and A. Waibel. Hierarchies of Neural Networks for Connectionist Speech Recognition. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN'98)*, April 1998.

[16] R. Gross. Run-on Recognition in an On-line Handwriting Recognition System. Master's thesis, Universität Karlsruhe, June 1997.

[17] R. Haeb-Umbach and Ney H. Improvements in Time Synchronous Beam Search for 10000-Word Continuous Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 2:353–356, 1994.

[18] W. Herrmann. *Lehrbuch der modernen koreanischen Sprache.* Helmut Buske Verlag, 1994.

[19] H.-W. Hon and K.-F. Lee. CMU Robust Vocabulary-Independent Speech Recognition System. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 889–892. IEEE, IEEE; New York, NY, May 1991.

[20] X.D. Huang and M.A. Jack. Semi-continuous hidden Markov models for speech signals. *Computer, Speech and Language*, 3(3):239–251, july 1989.

[21] M.-Y. Hwang and X. Huang. Subphonetic modeling with Markov states – Senone. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 33–36. IEEE, IEEE; New York, NY, March 1992.

[22] F. Jelinek. A Fast Sequential Decoding Algorithm Using a Stack. *IBM Journal of Research and Development*, 13(6):675–685, November 1969.

[23] F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, 1997.

[24] F. Jelinek, R.L. Mercer, and S. Roukos. *Advances in Speech Signal Processing*, chapter Principles of Lexical Language Modeling for Speech Recognition, pages 651–700. M. Dekker Publishers, New York, NY, 1991.

[25] T. Kemp, P. Geutner, M. Schmidt, B. Tornax, M. Weber, M. Westphal, and A. Waibel. The Interactive Systems Labs View4You Video Indexing System. In R.H. Mannell and J. Robert-Ribes, editors, *Proceedings of the International Conference on Spoken Language Processing, ICSLP 98*, volume 4, pages 1639–1942. Australian Speech Science and Technology Association, Incorporated, November 1998.

[26] P. Kenny. A* Admissible Heuristics for Rapid Lexical Access. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 682–692. IEEE, IEEE Comput. Soc. Press, Los Alamitos, CA, May 1991.

[27] D. Kiecza, T. Schultz, and A. Waibel. Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR. In *Proceedings of the*

International Conference on Speech Processing (ICSP'99), pages 323–327. Acoustical Society of Korea, August 1999.

[28] R. Kneser and H. Ney. Improved Backing-Off for M-Gram Language Modeling. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 181–184. IEEE Signal Processing Society, IEEE; New York, NY, May 1995.

[29] O.-W. Kwon, K. Hwang, and J. Park. Korean Large Vocabulary Continuous Speech Recognition Using Pseudomorpheme Units. In G. Prószéky, G. Németh, and J. Mándli, editors, *Proceedings of the Eurospeech 1999*, volume 1, pages 483–486. Scientific Society for Telecommunications, September 1999.

[30] L. Lamport. *Latex - A Document Preparation System.* Addison Wesley, second edition, 1994.

[31] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and Z. Puming. Janus III: Speech-to-Speech Translation in Multiple Languages. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 99–102. IEEE, IEEE Comput. Soc. Press, Los Alamitos, CA, April 1997.

[32] A. Lavie, A. Waibel, L. Levin, D. Gates, M. Gavalda, T. Zeppenfeld, P. Zhan, and O. Glickman. Translation of Conversational Speech With Janus-II. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP 96*, volume 4, pages 2375–2378. IEEE, IEEE, New York, NY, October 1996.

[33] H.-S. Lee, J. Park, and H.-R. Kim. An Implementation of Korean Spontaneous Speech Recognition System. In *Proceedings of the International Conference on Signal Processing Applications and Technology, ICSPAT 1996*, volume 2, pages 1801–1805, October 1996.

[34] K.-F. Lee. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: the SPHINX System.* PhD thesis, School of Computer Science, Carnegie Mellon University, May 1988.

[35] K.-F. Lee, S. Hayamizu, H.-W. Hon, C. Huang, J. Swartz, and R. Weide. Allophone Clustering for Continuous Speech Recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal*

Processing, volume 2, pages 749–752. IEEE, IEEE; New York, NY, April 1990.

[36] C.J. Leggetter and P.C. Woodland. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer, Speech and Language*, 9(2):171–185, April 1995.

[37] H. Ney, L. Welling, S. Ortmanns, K. Beulen, and F. Wessel. The RWTH Large Vocabulary Continuous Speech Recognition System. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 853–856. IEEE Signal Processing Society, IEEE; New York, NY, May 1998.

[38] H.J. Nock, M.J.F. Gales, and S.J. Young. A Comparative Study of Methods for Phonetic Decision-Tree State Clustering. In *Proceedings of the Eurospeech 97*, volume 1, pages 111–114. The University of Patras, Wire Communications Laboratory, September 1997.

[39] F.Y.T. Park. *Speaking Korean, Book I.* 1977.

[40] D.B. Paul. Algorithms for an Optimal A* Search and Linearizing the Search in the Stack Decoder. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 693–696. IEEE, IEEE Comput. Soc. Press, Los Alamitos, CA, May 1991.

[41] D.B. Paul. The Lincoln Tied Mixture HMM Continuous Speech Recogniser. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 329–332. IEEE, IEEE Comput. Soc. Press, Los Alamitos, CA, May 1991.

[42] L.R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, volume 77 of *2*, pages 257–286. IEEE, feb 1989.

[43] M. Ravishankar. *Efficient Algorithms for Speech Recognition.* PhD thesis, School of Computer Science, Carnegie Mellon University, May 1996.

[44] I. Rogina. *Parameterraumoptimierung für Diktiersysteme mit unbeschränktem Vokabular.* PhD thesis, Universität Karlsruhe, June 1997.

[45] R.W. Schafer and L.R. Rabiner. *Digital Representations of Speech Signals*, pages 49–64. In Waibel and Lee [56], 1990.

[46] T. Schultz et al. Language Independent and Language Adaptive Large Vocabulary Speech Recognition. In R.H. Mannell and J. Robert-Ribes, editors, *Proceedings of the International Conference on Spoken Language Processing, ICSLP 98*, volume 5, pages 1819–1822. Australian Speech Science and Technology Association, Incorporated, November 1998.

[47] T. Schultz and A. Waibel. Das Projekt GlobalPhone: Multilinguale Spracherkennung. In B. Schroder, W. Lenders, W. Hess, and T. Portele, editors, *Proceedings of the 4th Conference on Natural Language Processing-KONVENS-98. Computers, Linguistics, and Phonetics between Language and Speech*, pages 179–189. Peter Lang, Frankfurt am Main, Germany, October 1998.

[48] T. Schultz and A. Waibel. Development of Multilingual Acoustic Models in the GlobalPhone Project. In *Proceedings of the first Workshop on Text, Speech, and Dialogue (TSD)*, pages 311–316. Masaryk University, September 1998.

[49] T. Schultz and A. Waibel. Multilingual and Crosslingual Speech Recognition. In *Proceedings of the DARPA Broadcast News Workshop*, pages 259–262. IEEE, February 1998.

[50] T. Schultz and A. Waibel. Experiments towards a Multi-language LVCSR Interface. In *Second International Conference on Multimodal Interfaces (ICMI '99)*, January 1999.

[51] T. Schultz and A. Waibel. Language Adaptive LVCSR Through Polyphone Decision Tree Specialization. In *Proceedings of the Workshop on Multi-lingual Interoperability in Speech Technology (MIST '99)*, September 1999.

[52] T. Schultz, M. Westphal, and A. Waibel. The GlobalPhone Project: Multilingual LVCSR with JANUS-3. In *Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop*, pages 20–27, April 1997.

[53] S.-C. Song. *201 Korean Verbs - fully conjugated in all the forms*. Barron's Educational Series, Inc., 1988.

[54] S. Umesh, L. Cohen, and D. Nelson. Frequency-Warping and Speaker-Normalization. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 983–986. IEEE Signal Processing Society, IEEE Comput. Society Press; Los Alamitos, CA, April 1997.

[55] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. Meeting Browser: Tracking and Summarising Meetings. In *Proceedings of the DARPA Broadcast News Workshop*, 1998.

[56] A. Waibel and K.-F. Lee, editors. *Readings in Speech Recognition.* Morgan Kaufmann, 1990.

[57] M. Westphal. Dimensionalitätsreduktion von Sprachsignalen mit statistischen und neuronalen Methoden. Master's thesis, Universität Karlsruhe, January 1994.

[58] M. Westphal, T. Schultz, and A. Waibel. Linear Discriminant – A New Criterion for Speaker Normalization. In R.H. Mannell and J. Robert-Ribes, editors, *Proceedings of the International Conference on Spoken Language Processing, ICSLP 98*, volume 3, pages 827–830. Australian Speech Science and Technology Association, Incorporated, November 1998.

[59] M. Westphal and A. Waibel. Towards Spontaneous Speech Recognition for On-Board Car Navigation and Information Systems. In G. Prószéky, G. Németh, and J. Mándli, editors, *Proceedings of the Eurospeech 1999*, volume 5, pages 1955–1958. Scientific Society for Telecommunications, September 1999.

[60] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev, and S.J. Young. The 1994 HTK Large Vocabulary Speech Recognition System. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 73–76. IEEE Signal Processing Society, IEEE; New York, NY, May 1995.

[61] M. Woszczyna. *Fast Speaker Independent Large Vocabulary Continuous Speech Recognition.* PhD thesis, Universität Karlsruhe, February 1998.

[62] M. Woszczyna and M. Finke. Minimizing Search Errors due to Delayed Bigrams in Real-Time Speech Recognition Systems. In *Proceedings of the*

International Conference on Acoustics, Speech and Signal Processing, volume 1, pages 137–140. IEEE, IEEE; New York, NY, May 1996.

[63] M. Woszczyna, M. Finke, T. Kemp, A. McNair, A. Lavie, L. Mayfield, M. Maier, I. Rogina, T. Sloboda, A. Waibel, P. Zahn, and T. Zeppenfeld. Janus II - Translation of Spontaneous Conversational Speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 409–412. IEEE, IEEE; New York, NY, May 1996.

[64] S.J. Young. The General Use of Tying in Phoneme-Based HMM Speech Recognisers. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 569–572. IEEE, IEEE; New York, NY, March 1992.

[65] S.J. Young. A Review of Large Vocabulary Continuous Speech. *IEEE Signal Processing Magazine*, 13(5):45–57, September 1996.

[66] H. Yu, M. Finke, and A. Waibel. Progress in Automatic Meeting Transcription. In G. Prószéky, G. Németh, and J. Mándli, editors, *Proceedings of the Eurospeech 1999*, volume 2, pages 695–698. Scientific Society for Telecommunications, September 1999.

[67] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal, and A. Waibel. Recognition of Conversational Telephone Speech Using the Janus Speech Engine. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1815–1818. IEEE, IEEE Comput. Soc. Press; Los Alamitos, CA, April 1997.

[68] P. Zhan, K. Ries, M. Gavalda, D. Gates, A. Lavie, and A. Waibel. Janus-II: Towards Spontaneous Spanish Speech Recognition. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP 96*, volume 4, pages 2285–2288. IEEE, IEEE, New York, NY, October 1996.

[69] P. Zhan and A. Waibel. Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition. Technical Report CMU-CMT-97-150, Carnegie Mellon University, Center for Machine Translation, May 1997.

[70] P. Zhan and M. Westphal. Speaker Normalization Based on Frequency Warping. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 1039–1042. IEEE Signal Processing Society, IEEE Comput. Society Press; Los Alamitos, CA, 1997 April.

[71] P. Zhan, M. Westphal, M. Finke, and A. Waibel. Speaker Normalization and Speaker Adaptation – A Combination for Conversational Speech Recognition. In *Proceedings of the Eurospeech 97*, volume 4, pages 2087–2090. The University of Patras, Wire Communications Laboratory, September 1997.

[72] World Wide Web:

```
http://isl.ira.uka.de/~tanja/gp/index.html
http://www.unicode.org/
http://www.chosun.com/w21data/html/news/
ftp://ftp.kaist.ac.kr/hangul/code/hcode/
http://pantheon.yale.edu/~jshin/faq/qa8.html:
   Contains a description of the hangul
   code translation tool hcode.
```