AUTOMATIC RECOGNITION OF KOREAN BROADCAST NEWS SPEECH

Ha-Jin Yu, Hoon Kim, Jae-Seung Choi, Joon-Mo Hong,
Kew-Suh Park, Jong-Seok Lee, and Hee-Youn Lee

Information Technology Lab. MI group
LG Corporate Institute of Technology
hajin@lgcit.com

ABSTRACT

This paper describes preliminary results of automatic recognition of Korean broadcast-news speech. We have been working on flexible vocabulary isolated-word speech recognition, and the same HMM models are used for broadcast-news continuous speech recognition. The recognizer is trained by using phonetically balanced isolated words speech, rather than the broadcast news speech itself. In this research, we use several different lexica to investigate the recognition performance according to the length of the words. We also propose a long-distance bigram language model, which can be used at the first stage of the search, so that it can reduce the recognition errors caused by earlier pruning of correct hypothesis.

1. INTRODUCTION

Recently, broadcast news speech[1] has been a popular material for developing large vocabulary continuous speech recognizer, because of its richness of problems to be solved such as spontaneous speech, degraded speech, and so on. This paper describes preliminary results of Korean broadcast news speech recognition to reveal some problems and possibilities of Korean speech recognition.

We have to solve several problems to build a large vocabulary continuous Korean speech recognizer. One of the problems is that Korean words have a large amount of inflections. A verb can have about 2,000 inflections maximum, so if there are only 1,000 verbs, than we must have 2 millions of distinct words in a lexicon. There has not yet been a good solution published to deal with this problem, and only several thousand words have been used in Korean continuous speech recognition experiments so far. To extend our research, we have to decide the entry words of the lexica. As a beginning, we investigate the effect of entry words by using three lexica with entry words of different sizes.

We also propose a long-distance language model which can be used at the first pass of the search algorithms. Since correct hypothesis, pruned at earlier stages, can not be recovered at later steps, it is very important to use superior knowledge at the first pass. Language models which can represent knowledge about longer history have been proposed [4][5][6], but they are used at later passes because they need histories of many paths to be used in time-synchronous Viterbi beam search. In this research we propose a way of using such knowledge at the first pass, by checking existence of words in the word lattice history regardless of the paths.

This paper is organized as follows. In the next section, we describe the corpus for training and testing the system and we show the flow of the system. In Section 3, we explain the language model we propose. In Section 4, we present the experimental results. Finally, in Section 5, we give summary of this research.

2. THE CORPUS AND AN OVERVIEW OF THE SYSTEM

2.1. The Corpus

For training data, we use 6,700 phonetically balanced noun words spoken by 240 males and 160 females, so that the system is gender-independent. The total number of utterances is about 40,000. We use only noun words for training because Korean verbs and adjectives have the same ending sounds for almost all words. The system is tested with broadcast TV news speech collected over nine days (about six hours), and the weather forecast speech collected over 65 days. The news data consist of 3,516 sentences, and the weather forecast data consist of 857 sentences. The number of distinct words of the data is 17,671. The test data include sentences with maximum number of words 70, and the sentences are uttered at a rate of 21 phonemes per second on average. The database is divided into the sets shown in table 1, according to the speakers and the noise conditions. The utterances are manually segmented into sentences.

2.2. System Configuration

The training data and broadcast TV news test speech are sampled at 16 kHz 16 bit, and a Hamming window with a

width of 20 ms is applied at every 10 ms. Then a set of 12 LPC-derived cepstral coefficients and their derivatives are computed. We use 4 feature streams – 12 cepstra, 24 delta cepstra, 12 delta delta cepstra, and power including its first and second derivatives.

We use shared-state context dependent phoneme SCHMMs. Each left-to-right HMM model has three states. We compose the models of unseen triphones based on triphone similarities measured by comparing the left and right contexts[2].

Our search procedure consists of three passes[3] as shown in Figure 1. In the first pass, we use time-synchronous Viterbi beam search using bigram language model, which is derived from transcribed broadcast-news text collected over a year. Then in the second pass, we use A* backward search to find N-best sentence candidates by using bigram. In the first and the second pass, we also use the language model we propose in the next section. In the last pass, the top N sentences are re-ordered by using a more detailed language model such as trigram.

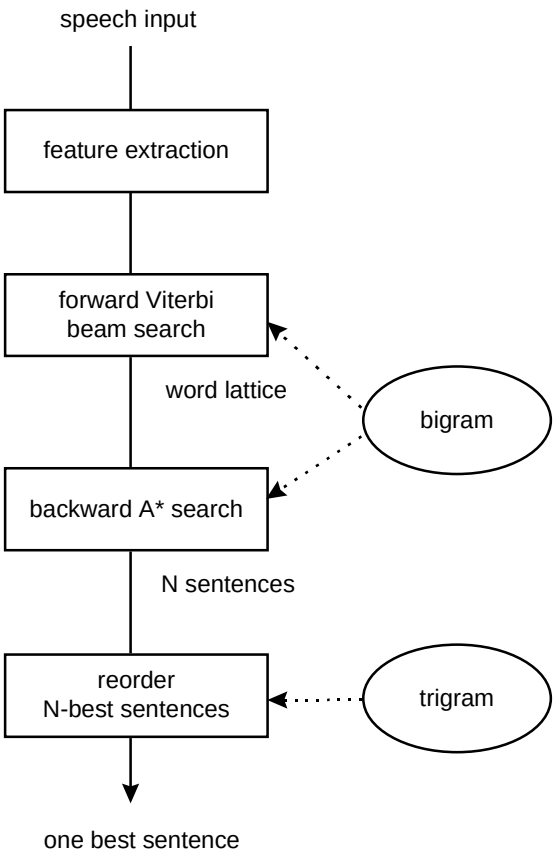

Figure 1: Search procedure consists of three passes

speakers	gender	number of sentences
anchors	male	466
	female	399
reporters	male	1908
	female	255
interviewees	male	425
	female	63
total		3516

Table 1: The corpus are segmented into the sets according to the speakers and the noise conditions

3. PROPOSED LONG-DISTANCE LANGUAGE MODEL

3.1. Background

Word bigram and trigram are the most commonly used language models. They are simple and easy to implement, but they only represent the relationships of neighboring words. If a sentence begins with a word which has a meaning of past tense, but ends with a word of present tense, we can conclude that the sentence may be grammatically wrong. Also, if a sentence begins with 'if', then we can expect that there is a high possibility of the presence of 'then' in the sentence afterward. However, the short-term conventional N-grams can not have enough information for them.

There have been attempts to capture some of the information present in the longer-distance history, such as trigger pair [4] and extended bigrams [5]. The long-distance language models are used mainly in the later passes such as stack decoder or N-best reordering stages. Any language model that looks further back into the history is harder to incorporate in the first time-synchronous Viterbi beam search, because multiple words paths that end in with same word are usually merged and all but one of the paths are lost. However, applying superior knowledge may be too late at some points, since the correct hypothesis may have already been pruned[4].

3.2. Proposed Language Model

In this research, we propose a long-distance language model which can be used at the first pass, so that the detailed knowledge can be used at an earlier stages, consequently reducing the unrecoverable errors.

A bigram probability of w_j given a one word history w_i

can be defined as $P(w_j|w_i)$, where the two words w_i and w_j are adjacent words (j=i+1) in a sentence. In the language model proposed in this research, the two words are not defined as adjacent words. Let us define forward and backward long-distance bigram probability of w_j given forward and backward one word histories w_i as

$P_{forward}(w_j \mid w_i)$, $i < j-1$,

$P_{backward}(w_j \mid w_i)$, $i > j+1$,

respectively, in a sentence consisting of words $w_1,w_2,...,w_n$. The forward probability $P_{forward}(w_j \mid w_i)$ is used in time synchronous forward Viterbi beam search, which is the first pass in our system. Here, w_i can be a word in the word lattice that is hypothesized before the time when w_j is being hypothesized. The probability is applied at the same time when a bigram $P(w_j|w_k)$ is applied, where w_k is a word hypothesized after w_i. By using any word in the word lattice, we do not have to keep all the paths as we have to when applying n-grams. The backward probability $P_{backward}(w_j \mid w_i)$ is used in time asynchronous A* backward search. In this case, w_i is a word which is found before w_j in a backward path, that is, a word after w_j in a hypothesized sentence.

By using this long-distance language models, we can use more knowledge at early stages, so that the correct hypothesis may not be pruned.

4. EXPERIMENTS

First, we tested the system by using several different lexica of different sizes. Since Korean sentences are composed of ***eojeol***s (which are similar to ***bunsets*** in Japanese) rather than words, the structure of the lexicon can affect the recognition performance and the system size. The lexicon can be composed of the ***eojeol***s or smaller units such as bases and affixes, which are the constituents of ***eojeol***s. If we use the ***eojeol***s as entry words in the lexicon, the nominal number of distinct words is increased because of the inflection. An ideal solution is to build a lexicon with only basic form of each word, and apply inflection rules in the search, but the systematic way of the solution has not yet been found. We can also use morphemes as basic units, but then the performance goes down because many words will consist of less than three phoneme-like units. Moreover, in Korean, the morphemes change in very complicated ways when they are assembled into words.

In this experiment, we first tested with only the weather forecast section of the broadcast news to compare the performances of the system with different lexica. We build three lexica with different vocabulary sizes. The first lexicon consists of ***eojeol***s (table 2(3)) and another consists of units as small as morphemes (table 2(1)), and the other consists of units with the sizes between ***eojeol***s and morphemes (table 2(2)). We tested the system with the tree lexica, without applying the language model we have proposed. As shown in table 2 (1)~(3), the tree lexica consist of 972, 1496, and 2288 words each, and we can observe that the performance is degraded as we subdivide the units of the recognition. The reason of the lower performance when we use the small units is due to the small number of phonemes in each word. Therefore, we can expect that the performance may be increased if we use cross-word PLU models.

On the other hand, the weather forecast news can be divided into two parts, one with vocabularies used frequently for weather forecast only, and the other with those which are not directly related to forecast such as season's greetings, or health care in different weathers. If we test the system with only the former part, we can get the high performance as shown in table 2 (4).

To examine the performance of the proposed long-distance language model, we fix the number of active states at each frame in the first time-synchronous Viterbi beam search path to 1,000, and applied the forward probability explained in previous section. We also applied the backward probability to the time-asynchronous A* backward search. As a result, the word recognition error rate decreases by 8.8%, from 16.3% to 14.8%. The recognition time is reduced by 10.6% from 11.1 minutes to 10.5 minutes per sentence on average in a Ultra Sparc workstation. Table 3 shows the overall recognition rate of the news speech database.

	(1)	(2)	(3)	(4)
number of words	972	1496	2288	358
number of sentences	857	857	857	213
word recognition rate	37.0 %	49.9 %	84.2 %	89.3 %
sentence recognition rate	11.1 %	15.9 %	59.2 %	64.3 %

Table 2: Recognition rate according to the lexica

speakers	gender	word recognition rate
anchors	male	87.7 %
	female	90.2 %
reporters	male	56.1 %
	female	79.9 %
interviewees	male	23.1 %
	female	27.3 %
total		59.9 %

Table 3: Recognition result of the broadcast news speech according to the data sets

5. CONCLUSIONS

This research is a preliminary experiment for large vocabulary Korean broadcast news speech recognition. We use phonetically balanced isolated words to train the system instead of using the broadcast news speech corpus itself. When we investigate the system with three lexica of different sizes of the basic units, the system performance varies from 49 % to 81 % in word accuracy as the size of the vocabularies varies from 1,000 words to 2,300 words. We can get higher performance when we use ***eojeol***s as the basic units. We can also reduce the word recognition error rate by 8.8%, and the recognition time by 10.6%, when we use the forward-backward long-distance language model we propose.

In the experiments, we can show that a properly trained set of context-dependent Korean phoneme models can be used in any continuous speech recognition applications when appropriate language models are provided.

6. REFERENCES

1. P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk, E.W.D. Whittaker & S.J. Young, "The 1997 HTK broadcast news transcription," Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, Virginia.

2. M. Y. Hwang, X. D. Huang, F. A. Alleva, "Predicting Unseen Triphones with Senones," IEEE Trans. Speech and Audio Processing, vol. 4, no. 6, pp. 412-419, Nov. 1996.

3. R. Schwartz, Y.L. Chow, "The N-Best Algorithm: An efficient and exact procedure for finding the N most likely sentence hypotheses," Proc. of ICASSP, pp.81-84, 1990Lyon, R.F., and Mead, C. “An Analog Electronic Cochlea,” *IEEE Trans. ASSP* 36: 1119–1134, 1988.

4. Rosenfeld, R. Adaptive Statistical Language Modeling: A Maximum Entropy Approach. Ph.D. thesis, School of Computer Science, Carnegie Mellon University, 1994.

5. Niesler, T. Category-based statistical language models, Ph.D. thesis, St. John*s College, 1997.

6. J.H.Wright, G.J.F.Jones and H.Lloyd-Thomas, A consolidated language model for speech recognition, Proc. of Eurospeech 93, Berlin, vol. 2, pp. 977-980, 1993.