THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.20, No.2E 2001. 6. pp. 3~8

Korean LVCSR for Broadcast News Speech

Gang Seong Lee*

*Computer Engineering Dept. KwangWoon Univ.
(Received 22 November 1999 ; revised 4 January 2001 ; accepted 9 April 2001)

Abstract

In this paper, we will examine a Korean large vocabulary continuous speech recognition (LVCSR) system for broadcast news speech. The combined vowel and implosive unit is included in a phone set together with other short phone units in order to obtain a longer unit acoustic model. The effect of this unit is compared with conventional phone units. The dictionary units for language processing are automatically extracted from eojeols appearing in transcriptions. Triphone models are used for acoustic modeling and a trigram model is used for language modeling.

Among three major speaker groups in news broadcasts--anchors, journalists and people (those other than anchors or journalists, who are being interviewed), the speech of anchors and journalists, which has a lot of noise, was used for testing and recognition.

Keywords: *Speech recognition, LVCSR, Language modelling, Broadcast news*

I. Introduction

Lately, interest in large vocabulary continuous speech recognition research has been shifting away from read speech data to speech data found in the real world. Broadcast news speech over radio or TV is a good example[1,2]. It is from humans--not recorded for a machine--easy to record, and most of the time a correct transcription is provided. It is also spoken by many different people with different speaking styles. The speech data contains a wide variety of noise environments and channel conditions. Unlike other domain-specific targets, this is an almost open domain target, which makes language modeling difficult. For many reasons, this is a real challenge.

The database, KBN01 (Korea Broadcast News) set, currently contains 16 KBS 9PM evening news shows of one hour each (a total of 13.7 hours). Long, unsegmented speech data is processed into smaller segments. Each segment contains few sentences, and speaker ID and noise level information are tagged on each segment.

In this paper, a new acoustic unit is employed, a combination of vowels and implosives, together with conventional phone units. This is the phoneme sequence that changes acoustic features drastically depending on preceding or succeeding phonemes. The effectiveness of these units will be shown by comparing those with conventional phone units.

Since Korean is a highly inflected language, choosing dictionary units is a difficult problem in HMM based Korean LVCSR system. Using the compound units (eojeols) that agglutination process as dictionary units gives unmanageably large dictionaries with extremely high Out-of-Vocabulary (OOV) rates[3].

Korean words are only built from about 3500 different syllables, where each syllable consists of one to four

Corresponding author: Gang Seong Lee (gslee111@daisy.kwangwoon.ac.kr)
KwangWoon Univ., Wolgye-Dong Nowon-Gu Seoul, Korea 139-701

phonemes. Choosing these syllables as dictionary units provides small dictionaries and OOV-rates below one percent. It increases, however, the confusability in acoustic modeling and language modeling since it is too short.

An automatic dictionary unit detaching algorithm applied to the system is also presented, which is the basic unit of language modeling and a dictionary, and it is shown how effectively it divides compound words or inflected words into their own parts. This algorithm does not extract exactly the same unit as the morphological unit, but it reduces greatly the amount of time and effort.

The recognition engine is based on the JRTk (Janus Recognition Toolkit) which was developed at (CMU) Carnegie Mellon Univ. in the U.S. and at Karlsruhe Univ. in Germany.

II. System Description

2.1. Signal Processing

Each news show is converted from videotapes into MPEG file format with a sampling rate of 44.1kHz and then downsampled to 16kHz to get a PCM file format. The frame window size is 20msec, and frames are spaced at 10msec intervals. The system produces the following feature stream of 43 dimensions:

mel ceptrum of 13 order	(NMCEP)
delta melcepstrum	(DFEAT)
delta delta melcepstrum	(DDFEAT)
zero crossing rate	(ZERO)
logpower	(POWER)
delta logpower	(DPOWER)
delta delta logpower	(DDPOWER)

and then the dimension is cut to 24 by using K-L expansion[5].

2.2. Database

Each of the 16 entire news show segments is divided into smaller segments. For computational convenience, speech is segmented where the speaker is changed, where the subject is changed, where noise or music is present, or where it contains more than 2-3 long sentences. Each segment includes information on transcription, speaker ID, noise level, speaker group ID, and the type of speech. Phonetic labeling is done automatically during the training procedure using forced viterbi alignment.

There are three different speaker groups: ANC (anchors), JRN (journalists or reporters) and P (people interviewing). Four noise conditions are defined: clean speech (N0), speech with little noise (N1), speech with some noise (N2), speech with loud noise (N3), speech with very loud noise (N4). Four types of speech are included: speech, noise (of people or nature), music, and foreign language (English, Chinese, Japanese). The number of anchors is four (2 male / 2 female), the number of journalists is 191 (183 male / 8 female) and the number of people interviewed is 774 (male 633 / female 141). The total number of people involved in the news shows is 969 (818 male / 151 female).

2.3. Acoustic Modeling

All polyphones have 3-state left-to-right topology except silence, which has only one state. These sub-polyphones (states) are clustered to build 2,000 sub-allophones[4]. To make each model of the sub-allophones, a codebook containing 16 Gaussians with diagonal covariances is used.

2.4. The Phone Set

The conventional phone set is as follows:

/ K Kk N T Tt R M P Pp S Ss C Cc Ch Kh Th Ph H nc nh lm lh a ya yae eo e ae yeo ye o wa oe yo u weo we wae wi yu eu yi i n l m ng k kk ks t lk lp ls lth lph p ps s ss c ch kh th ph h /

Longer unit generally has better performance. The main reason a longer unit than a phoneme is not popular, such as a syllable or word, is that there are too many context-dependent units to be processed in a finite computer.

It doesn't have to be a logical unit like a syllable or word for a longer unit. Assume that /a b c/ is a phoneme sequence. Neighboring phonemes (/a b/, /b c/) affect each other. Phonemes /a/ and /c/ have a weaker influence on each other. In Korean, there is a certain sequence in which /c/ affects /a/ as strongly as it affects /b/. This happens when /a/ is a vowel and /b/ is an implosive.

For instance, /a k^/ + /i/ becomes /a: k i/, /a k^/ + /k

i/ becomes /a k^ k' i/ and /a k^/ + /m yeo ng/ becomes /a ng m yeo ng/, where /k^/ is an implosive consonant of /k/, /k' / is an emphasis consonant, /a:/ is a long vowel and /a/ is a short vowel. Making /a k^/ one phonetic unit, the most drastic changes can be absorbed in one unit, without significantly increasing the number of units. In Korean, there are 21 (composite) vowels and 47 (composite) consonant (20 for Chosung, 27 for Jongsung) symbols, although some consonants share the same pronunciation depending on the neighboring phonemes. In the system, we defined 26 consonants, 21 vowels and 410 vowel and implosive-like combinations. This number of combinations is generated theoretically. Thus, there are a lot of combinations not used in actual speech. The actual number of combined units occurring in training data is 95, which is much less than the theoretical number of 410. The total number of phonetic units used in the system becomes 142 plus Noise(1).

III. Dictionary Unit Determination

3.1. Forward-Backward Splitting Algorithm

The selection of a dictionary unit for language modeling is not a simple matter, especially in a language that is highly inflected. In a language such as English, the word level seems to be a useful abstraction. For Asian languages such as Korean and Japanese, the basic unit is chosen at a subword unit. In Korean, one verb is inflected into more than 30 different forms.

There are two ways to approach the acquisition of subword units from text corpus in Korean. One is having people divide tokens into morphological units. Although automatic morphological unit extraction based on the predefined morphological lexicon is possible, it can never be perfect because there are a lot of ambiguous cases that cannot be decided without prior semantic analysis. The other approach is automatic dictionary unit extraction from a list of eojeols that share common parts with other eojeols. This approach is simple and easy to implement, even though the result might be a little different from morphological units.

One disadvantage is that this unit or subword depends on the text corpus. If the text corpus is changed, then the unit will be possibly changed. This unit has no direct relationship with linguistic morphemes. Nevertheless, this approach is of great interest because it is fast, easy to implement, and requires no human interference. The algorithm used here is a simple automatic subword-generating method. It works forward and backward to produce stems and tails. The dividing unit is the syllable.

Prior to any processing, all eojeols or tokens that are separated by space in the text corpus are collected and sorted. Having word tokens, we set the sharing syllable number and counter variable with the length of syllables sharing other tokens and the consecutive number of words sharing the same part of the token. As an illustration, consider the following data structure (symbols are syllable):

	token no.	sharing syllable no	counter	token dic.
	:	:	:	:
→	5	2	2	a b a d e
	6	2	1	a b c
	7	1	0	a b d
	8	1	0	a c c
	9	2	1	a d d
	:	:	:	:

subword dic: a b a

new token dictionary: d e

minlen = 2

Assume that we are to process token 5. For each token, the following steps are applied:

Step 1) Check subword_dic for any registered subwords greater than the shared syllable. If found, take the longest one and go to Step 2; otherwise go to Step3).

ex) In the example, [a b a d e], [a b a d] and [a b a] are searched. [a b a] is found.

Step 2) Take out the registered part from the token and put the rest of it into a new token dictionary. Go to Step 5.

ex) Take out [a b a] from [a b a d e] and put [d e]

into a new token dictionary.

Step 3) If there is no sharing part or the number of shared syllables is less than minlen, then register whole word in the dictionary. Go to Step 5. Otherwise go to Step 4.

ex) Sharing syllable length of token 8 is 1 and minlen = 2, thus register [a c c].

Step 4) Register sharing part and put rest of it into new token dictionary for all tokens which share the common part and move token pointer to the last processed token.

ex) If you are on token 6, register [a b] and put [c] of token 6 and [d] of token 7 into new token dictionary and move token pointer to 7.

Step 5) Add token pointer 1

Step 6) Repeat Steps 1-5 until all the tokens are processed.

Step 7) Set token dictionary to a new token dictionary. If this is first or second episode of Step7, then reverse syllables in the new token dictionary. Recalculate sharing syllable no. and counter. Repeat Steps1-6.

ex) [d e] becomes [e d]. This is for extracting tails.

Step 8) Repeat Steps 1-7) until all the subwords are registered and the new token dictionary is empty.

* When the token dictionary is reversed by syllables and you register a subword in a dictionary, reverse syllable order should be considered in Step 3 and Step 4.

Here is an actual example.

Original text :

Pu-Chae-Ka Manh-eun Ki-eop-Kki-Ri Hap-Pyeong-Ha-Myeon Tto Ha-Na-yi Teong-Chi Kheun Pu-Sil-Ki-eop-Man Than-Saeng-Han-Ta-Neun Keos-i Ceong-Pu-yi Phan-Tan-ip-Ni-Ta

(부채가 많은 기업끼리 합병하면 또 하나의 덩치 큰 부실기업만 탄생한다는 것이 정부의 판단입니다)

Separated into subword units:

Pu-Chae Ka Manh-eun Ki-eop Kki-Ri Hap-Pyeong-Ha Myeon Tto Ha-Na yi Teong-Chi Kheun Pu-Sil-Ki-eop Man Than-Saeng-Han Ta-Neun Keos-i Ceong-Pu yi Phan-Tan-ip-Ni Ta

(부채가 많은 기업끼리 합병하면 또 하나의 덩치 큰 부실기업만 탄생한다는 것이 정부의 판단입니다)

Separated into morphological units:

Pu-Chae Ka Manh-eun Ki-eop Kki-Ri Hap-Pyeong Ha-Myeon Tto Ha-Na yi Teong-Chi Kheun Pu-Sil Ki-eop Man Than-Saeng Han-Ta-Neun Keos i Ceong-Pu yi Phan- Tan ip-Ni-Ta

(부채가 많은 기업끼리 합병하면 또 하나의 덩치 큰 부실기업만 탄생한다는 것이 정부의 판단입니다)

As you can see, although there are some differences, both results are quite similar. The unit 'Hap-Pyeong-Ha (합병하)' came from the largest sharing unit of 'Hap-Pyeong- Ha-Ko (합병하고)', 'Hap-Pyeong-Ha-Neun (합병하는)', 'Hap-Pyeong-Ha-Myeon (합병하면)' etc., the unit 'Than-Saeng-Han(탄생한)' from 'Than-Saeng-Han (탄생한)' and 'Than-Saeng-Han-Ta-Neun (탄생한다는)'.

3.2. Language Modeling

Text training material is collected from two major Korean broadcast companies - KBS, MBC. The total text from 10 months of transcription is used for language modeling. 1.54M tokens or 2.6M subword tokens were counted. The number of different tokens is 250K, and the vocabulary size of the subword is 45K. The subword split algorithm reduced the vocabulary size by the factor of 0.18. The number of bigrams and trigrams are 740K and 1.6M respectively. The OOV (Out-of-Vocabulary) rate is 2.8%. The Katz smoothing technique[6] is applied in trigram modeling. The counter number for Good-Turing estimate is 7.

IV. Experiments and Discussion

Fourteen (14) out of 16 news shows are used for training, and two are used for testing. One acoustic model is made for all anchors and journalists. Speech segments of all levels of noise (N0-N3) are involved for acoustic modeling, except N4 (very loud noise) segments. Speech recognition accuracy is conventionally expressed in terms of word error rate (WER). To calculate this, an alignment of the

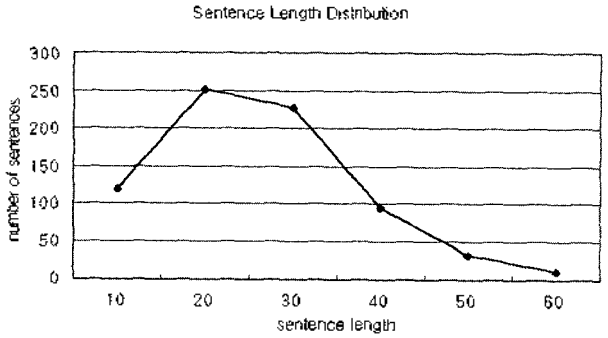

Figure 1. Sentence length distribution.

hypothesized and reference transcription is made and the number of insertion (I), deletion (D) and substitution (S) errors are counted. Word error is then expressed by:

$$WER = 100 \times \frac{S + D + I}{\#\_of\_spoken\_words}$$

The total number of test sentences is 735. The distribution of sentence length by 10 is shown in Fig 1.

The WER of anchors and journalists is 47.42% and 70.62% for Phone Set 1 (conventional phone set), 49.94% and 69.94% for Phone Set 2 (extended phone set) respectively. There is no prominent difference between the two phone sets. Although Phone Set 2 uses longer acoustic units partially, the number of states within those units is the same as the other set of 1 phone set units. The reason why the speech of anchors is recognized better than that of journalists is that, usually, the speech of anchors falls into noise levels ranging from N0-N2 and the speech of journalists is typically within N2-N4.

Table 1. WER by phone sets.

1 Phone set	47.42%	70.62%
2 Phone set	49.94%	69.94%

Table 2. WER by noise level.

N0~N1	53.89%	56.14%
N2	70.24%	70.49%
N3	75.47%	73.11%
N4	83.69%	81.94%

Table 3. WER by sentence length.

0~10	38.86	41.91
11~20	73.41	73.34
21~30	69.41	69.51
31~40	69.45	70.43
40~	69.81	66.14

The WER by noise level is illustrated in Table 2. As noise level increases, the WER increases. Particularly when the speech is clean, the WER (53.89%, 56.14%) is much lower than others. That means noise is the major factor that affects recognition performance. Here again, it is hard to find any difference between the two phone sets.

The performance according to sentence length is measured. The WER by length of sentence is depicted in Table 3. For very short sentences with 10 or fewer words, the WER is considerably lower (38.86%). Longer sentences, however, do not seem to exhibit any differences. The WER for sentences less than 11 subwords is much lower than others. The main reason for this is there are a lot of sentences that fit into typical sentence structures, like 'I'm XXX, KBS news'.

V. Conclusions

The system for broadcast news large vocabulary speech recognition is described. For the phone set, new combined units of vowels and implosives are added to the normal phoneme set to comprise drastically changing acoustic features. The result was, however, just as good as the conventional phone set. The dictionary unit extracting algorithm is presented. The less noise presented in speech, the better the performance it showed. With respect to the length of sentences, in shorter sentences, the error rate is low; but in longer sentences the results were indistinguishable. We got 47.42% WER from anchor's speech and 70.62% WER from journalists' noisy speech when conventional phone set is used for acoustic modeling.

Acknowlegdements

The research has been conducted by the Research Grant of Seoam Research and Scholarship Foundation in 1998.

References

1. Steven Wegmann, Puming Zhan, Larry Gillick, "Progress in Broadcast News Transcription at Dragon Systems," Proc. ICASSP-99, Phoenix, AR, May 1999.
2. Scott S Chen, Ellen M Eide, Mark J. F. Gales, Ramesh A Gopinath, Dimitri Kanevsky, Peder A Olsen, "Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News," Proc. ICASSP-99, Phoenix, AR, May 1999.
3. Lee, Hang-Seop and Park, Jun and Kim, Hoi-Rin, "An Implementation of Korean Spontaneous Speech Recognition System", *Proceedings of ICSSPAT96*, pp 1801-1805, Seoul, Korea, 1996.
4. Keinosuke Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press Inc. 1990.
5. Ivica Rogina, "Automatic Architecture Design by Likelihood-Based Context Clustering With Cross Validation", *Proceedings of Eurospeech-97*, 1997.
6. Stanley F. Chen, "Building Probabilistic Models for Natural Language," Ph.D. thesis, Harvard Univ, May 1996.

[Profile]

• Gang Seong Lee

Lee, Gang Seong was born on January 15, 1964 in Seoul, Korea.
He received the B.S. degree in computer engineering, and the M.E. and Dr. Eng. degrees in the same field from Kwangwoon University, Seoul, Korea, in 1986, 1988, and 1993, respectively. In 1991, he joined the faculty of Computer Engineering Dept. in Kwangwoon University, and he has been an Associate Professor since 1998. he studied at Carnegie Mellon University since Aug. 1998, for a year, as a visiting research scholar at School of Computer Science. His research interests are speech recognition and language modelling.