JAPANESE LARGE-VOCABULARY CONTINUOUS-SPEECH RECOGNITION USING A BUSINESS-NEWSPAPER CORPUS

Tatsuo Matsuoka[†], *Katsutoshi Ohtsuki*[††], *Takeshi Mori*[†††], *Sadaoki Furui*[†],[†††], *and Katsuhiko Shirai*[††]

[†]NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo 180, JAPAN
[††]Waseda University, [†††]Tokyo Institute of Technology

ABSTRACT

We studied Japanese large-vocabulary continuous-speech recognition (LV CSR) for a Japanese business newspaper. To enable word N-grams to be used, sentences were first segmented into words (morphemes) using a morphological analyzer. Newspaper articles for about five years were used to train N-gram language models. To evaluate our recognition system, we recorded speech data for sentences from another set of articles. Using the speech corpus, LV CSR experiments were conducted. For 7k vocabulary, the word error rate was 82.8% when no grammar and context-independent acoustic models were used. This improved to 20.0% when both bigram language models and context-dependent acoustic models were used.

1. INTRODUCTION

Large-vocabulary continuous-speech recognition (LV CSR) is being extensively studied for American/British English, French, German, and Italian languages for business newspapers such as the Wall Street Journal [1-9]. However, no research efforts have been reported for Japanese. This is mainly because Japanese sentences are written without spaces between words, so they are very difficult to segment automatically. Therefore, it is difficult to estimate a word N-gram language model, which is very helpful for LV CSR. There has been no speech corpus for Japanese LV CSR for business newspapers. Therefore, we first designed a speech corpus that could be used for Japanese LV CSR research.

We designed a business-newspaper speech corpus using the Nihon Keizai Shimbun (Nikkei newspaper). A word frequency list of 623k words was derived from 6.8M sentences. Vocabulary sizes of 7k, 30k, and 150k were defined to have the same coverage as vocabulary sizes (5k, 20k, and 64k) in the WSJ task.

We studied acoustic modeling and language modeling for Japanese LV CSR using this speech corpus. In this paper, we report on the design of the speech corpus and the effects of context-dependent acoustic modeling and word N-gram language modeling.

2. DESIGN OF THE CORPUS

Newspaper articles for five years were divided into two parts: four years and nine months for training and three months for testing.

2.1 Text preprocessing

Texts were preprocessed before morphological analysis. This was done to keep the sentences easy to read and also to keep morphological analysis correct so that language models could easily be estimated. Since our target is LV CSR and not sentence dictation, we discarded marks that are usually not pronounced in spoken communications.

Sentences that were too long were discarded from the training and test sets, because long sentences are not easy to read. We assumed the distribution of the sentence length, in terms of the number of words, was a normal distribution. The number of words in a sentence was 25.6 on average for both the training and test sets, and the standard deviations were 13.8 words for the training set and 13.5 words for the test set. Sentences with lengths in the range of the mean$\pm 2\sigma$ in terms of the number of words in a sentence were used to make a word frequency list and for training N-gram language models. The test set consisted of sentences with lengths in the range of the mean$\pm\sigma$ to ensure readability.

After these text preprocessing, there were 6.8M sentences and 180M words in the training set, and 342k sentences and 9.8M words in the testing set.

2.2 Morphological analysis

The segmentation into words requires sophisticated morphological analysis. Our morphological analyzer has a dictionary of 250k morphemes. The accuracy of the morphological analysis is about 95% for the Nikkei newspaper. In this study, we defined words by morphemes according to the lexicon of our morphological analyzer.

A word frequency list is a frequency-based-sorted list of words appearing in the training set. We had a list of 623k words. Since the morphological analyzer has a dictionary of 250k words, 373k out of 623k words were analyzed as unknown words. Most of the unknown words were proper nouns or very special technical terms.

To eliminate morphological errors and typographical errors, the sentences including words that did not appear in the top 150k words (coverage 99.6%) in the word frequency list were discarded.

2.3 Design of the corpus

Table 1 shows the coverage of our task and the LV CSR tasks for other languages [9]. There are more distinct words for

Table 1 Comparison of lexica and LM training corpora for different languages

	Nikkei (Japanese)	WSJ (English)	Le Monde (French)	Frankfurter Randschau (German)	Sole 24 (Italian)
Training text size	180 M	37.2 M	37.7 M	36 M	25.7 M
Distinct words	623 k	165 k	280 k	650 k	200 k
5k coverage	88.0 %	90.6 %	85.2 %	82.9 %	88.3 %
7k coverage	90.3 %	-	-	-	-
20k coverage	96.2 %	97.5 %	94.7 %	90.0 %	96.3 %
30k coverage	97.5 %	-	-	-	-
40k coverage	98.2 %	99.2 %	97.6 %	-	98.8 %
65k coverage	99.0 %	99.6 %	98.3 %	95.1 %	99.0 %
20k OOV rate	3.8 %	2.5 %	5.3 %	10.0 %	3.7 %

Table 2 Description of subsets

Subset	Description
7k	Sentences composed solely from 7k vocabulary
7k+	Sentences composed from 7k vocabulary and up to two OOV words
30k	Sentences composed solely from 30k vocabulary
30k+	Sentence composed from 30k vocabulary and up to two OOV words
30k++	Sentences composed from 30k vocabulary and more than two OOV words

Japanese and German than for the other three languages. Compound-words result in a large number of distinct words. In addition, inflection increases the nominal number of distinct words in Japanese. The out-of-vocabulary rate for Japanese is higher than that for English but lower than that for German. We defined 7k and 30k vocabularies to have the same coverage as the 5k and 20k vocabularies for the WSJ task.

We investigated the number of homophones[9]. Our Nikkei task had a 20.0% homophone rate in the lexicon. When the homophone rate in the lexicon is high, it is more difficult to perform LV CSR for that language, because homophones require some semantic knowledge to be understood.

To evaluate an LV CSR system, five subsets were defined according to the vocabulary size and the number of out-of-vocabulary (OOV) words in a sentence for each of the training and testing sets. Table 2 lists the description of subsets. The OOV words were limited to those appearing in the top 150k vocabulary, that is, our task was limited to 150k words.

Each of 54 speakers uttered 100 sentences, i.e., 20 from each of the subsets. Fifty sentences were selected from the training set, and fifty sentences were selected from the testing set. Speech was recorded simultaneously through a head-mounted Sennheiser (HMD-410) microphone and a desk-mounted Crown (PCC-160 phase coherent cardioid) microphone. Speech recorded through the head-mounted microphone was used for the experiments described in this paper.

The average number of words in a sentence for each subset ranged from 20.9 to 27.4, and the average time duration of a sentence ranged from 6.3 s to 8.6 s. As the number of vocabulary words increased the length increased, although not significantly.

3. ACOUSTIC MODELING

We evaluated context-independent, and diphone/triphone-context-dependent models. Each model was trained using a large population of speakers. Speech data from tasks other than the newspaper articles were used to train the acoustic models. In addition, the microphones used for recording the training and testing speech were different; training speech was recorded through desk-mounted microphones and testing speech was recorded through head-mounted microphones. Therefore, our task may be more difficult than the WSJ task, where the training and testing speech were recorded from the same task. We used phonetically-balanced-sentence speech to train the acoustic models. In total, more than 15,000 utterances from 58 speakers were used.

Every phone model has three states, except for the silence model, which has one state. Each state has four mixture Gaussian distributions. Speech was sampled at 12 kHz and digitized into 16 bits. The acoustic features used were 16 LPC derived cepstra and log-energy, and their first derivatives (delta-features).

Table 3 lists the results of the phoneme recognition experiments. The number of phoneme classes was 42 including silence. In these experiments, continuous speech recognition with a no-grammar network was carried out defining the phoneme as the recognition unit. Accuracy was calculated as

$$Accuracy = \left(1 - \frac{S + D + I}{N}\right) \cdot 100,$$

where S, D and I are the number of substitution, deletion, and insertion errors, respectively. Di2000 means diphone-context-dependent models whose training samples could be observed more than 2000 times in the training set. Tri*Number* means triphone-context-dependent models. The highest accuracy of 61.6% was achieved when Di700 and Tri300 were used with context-independent (CI) models with smoothing,.

Table 3 Phoneme recognition accuracy (%); Effect of context-dependent modeling

	CI	Di2000	Di1000	Di700	Di500	Di300	Di100
CI	49.2	58.0	58.4	58.2	57.9	57.2	56.7
Tri600	58.4	60.4	60.6	60.6	57.9	60.1	-
Tri500	59.4	61.0	60.9	61.0	60.5	60.6	-
Tri400	58.9	61.0	60.9	61.2	61.1	60.8	-
Tri300	60.6	61.5	61.2	***61.6***	61.5	61.3	-
Tri200	60.4	60.9	60.6	60.9	60.9	60.8	-
Tri100	60.9	61.3	60.8	61.0	61.1	60.9	-
Tri50	60.9	-	-	-	-	-	-

Table 4 Number of N-grams in the training set (upper) and average number of occurrences of each N-gram (lower)

	Unigram	Bigram	Trigram
7 k	7000	2.1 M	17.1 M
	24388.0	65.1	7.2
30 k	30000	4.9 M	30.5 M
	6121.9	33.8	5.1

4. LANGUAGE MODELING

Table 4 lists the distinct number of unigrams, bigrams, and trigrams observed in the training set, and the average number of occurrences of each corresponding N-gram. Most of the bigrams and trigrams are singletons (appeared only once in the entire training set). Since the average occurrences of bigram and trigrams are low, that is, we had only five to seven occurrences for each trigram on average, so the language models obviously needed to be smoothed. We used the back-off smoothing method proposed by Katz [10].

Using the smoothed N-gram language models, we evaluated test-set perplexity. All of the newspaper articles from the test set were used to calculate test-set perplexity. Table 5 shows the test-set perplexity for the Nikkei task compared with the WSJ task [8].

When the language models for the Nikkei task were estimated, punctuation marks were considered. Therefore, it is reasonable to compare the perplexities with the VP case of the WSJ task. However, quotation marks were omitted in our text preprocessing, and the definitions of words were different for our task and the WSJ task. Neglecting these differences in detail, the perplexities were fairly similar.

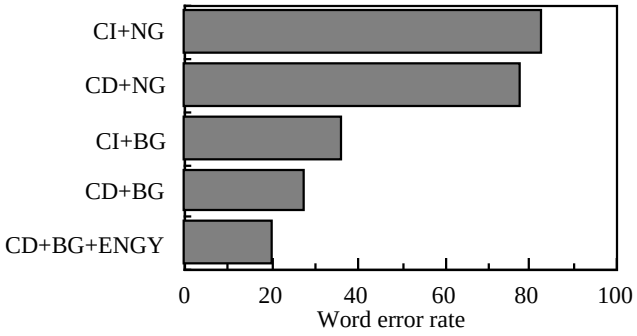

Figure 1 LV CSR experimantal results

5. CSR EXPERIMENTS

Continuous-speech recognition experiments were carried out for the 7k vocabulary task using the first 10 speakers' speech from the recorded speech corpus. Context-independent and intra-word context-dependent acoustic models were used. As context-dependent models, we used the model set that achieved the best result in the preliminary phoneme recognition experiments. Figure 1 illustrates the LV CSR results. The word error rate was obtained as

$$ErrorRate = \frac{S + D + I}{N} \cdot 100,$$

where *S*, *D* and *I* are the number of substitutions, deletions, and insertions, respectively [11].

The word error rate for the baseline system with context-independent phoneme models and no-grammar language models was 82.8% for the test set. This improved to 36.3% when the bigram language models were used. Using context-dependent phoneme models and introducing log-energy and Δ-log-energy further improved to 20.0%. The error rate was then approximately halved when bigram language models were incorporated. The further addition of the context-dependent acoustic models with log-energy and Δ-log-energy again reduced the remaining error by approximately half.

As the performance differed with the speakers, Figure 2

Table 5 Test-set perplexity

Nikkei			WSJ			
Vocabulary size	Language model	Test-set perplexity	Vocabulary size	Language model	Test-set perplexity	
					VP	NVP
7 k	Unigram	597	5 k	Unigram	-	-
	Bigram	82		Bigram	80	118
	Trigram	38		Trigram	44	68
30 k	Unigram	693	20 k	Unigram	-	-
	Bigram	124		Bigram	158	236
	Trigram	64		Trigram	101	155

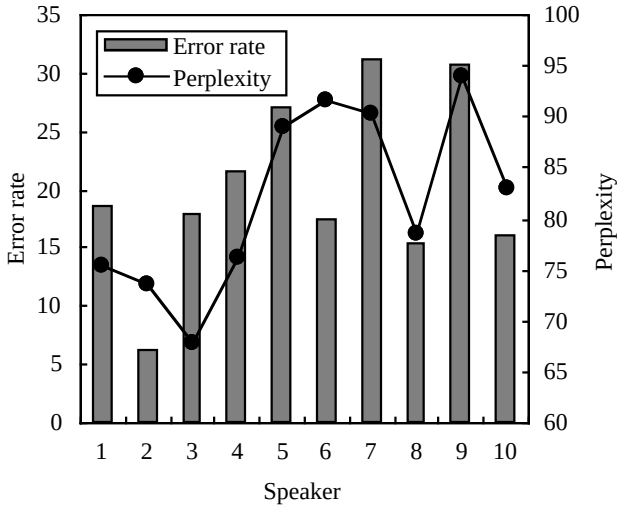

Figure 2 Word error rate and test-set perplexity for each speaker

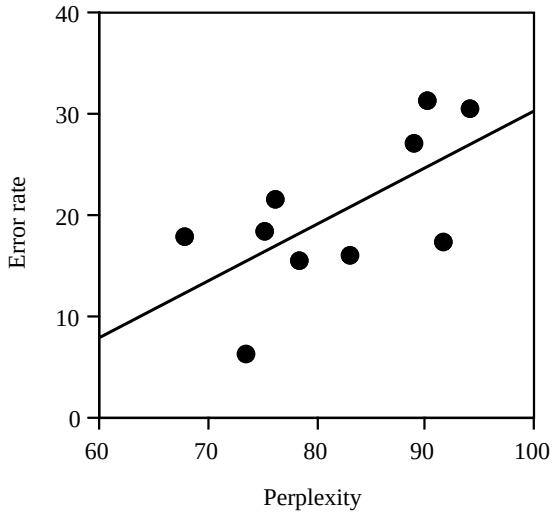

Figure 3 Relation between perplexity and word error rate

illustrates the results for each speaker. Each speaker read different sentences, which varied in recognition difficulty, so the difference in performance could be attributed to these differences. We calculated the perplexity for each speaker assuming that the sentences for one speaker made up a test-set. As shown in Figure 2, the test-set perplexity varied from 67.8 to 94.0. Figure 3 illustrates the relationship between the test-set perplexity and the word error rate. The results show that the error rate increased as the perplexities increased. The solid line indicates the first-order regression. The deviation from the line can be interpreted as variability due to speaker-dependent acoustical characteristics.

6. SUMMARY

We have described the design of a speech corpus for a Japanese business newspaper for LV CSR and have evaluated an LV CSR system.

Vocabulary sizes of 7k and 30k were defined according to word frequency, and sentences for the speech corpus were chosen from Japanese business newspaper articles covering five years. Fifty-four speakers contributed to the speech corpus.

An LV CSR system was evaluated using the first 10 speakers from the speech corpus. For the 7k vocabulary task, the word error rate for the baseline system, which used context-independent acoustic models and no-grammar language models, was 82.8%. This improved to 20.0% when context-dependent acoustic models and bigram language models were used. The error reduction, therefore, was 76%. Both bigram language models and context-dependent acoustic models were very effective in reducing the error rate. The bigram language models reduced the error by half. Similarly, the further addition of the context-dependent acoustic models again halved the remaining error.

We are further improving both acoustic models and language models. For the acoustic models, inter-word context-dependent models are currently being introduced instead of intra-word context-dependent models [12]. And for the language models, we are introducing trigram models in addition to bigram and unigram models.

ACKNOWLEDGMENT

We would like to thank Mr. Kazuo Tanaka of NTT Human Interface Laboratories for providing us with the morphological analyzer. We are also grateful to Nihon Keizai Shimbun Incorporated for allowing us to use the newspaper text database (Nikkei CD-ROM 90-94) for our research. We also thank the volunteer speakers who participated in the speech corpus collection.

REFERENCES

1. D. B. Paul, et al., Proc. ICSLP-92, pp. 899-902
2. Gauvain, et al., Proc. ICSLP-90, pp. 1097-1100
3. T. Robinson, et al., Proc. ICASSP-95, pp. 81-84
4. H. J. M. Steeneken, et al., Proc. Eurospeech-95, pp. 1271-1274
5. P. C. Woodland, et al., Proc. ICASSP-95, pp. 73-76
6. D. Pye, et al., Proc. Eurospeech-95, pp. 181-184
7. L. Lamel, et al., Proc. Eurospeech-95, pp. 185-188
8. D. B. Paul, et al., Proc. ICASSP-93, pp. 660-663
9. L. Lamel, et al., Proc. IEEE Automatic Speech Recognition Workshop, pp. 51-54, Snowbird, Dec. 1995
10. S. M. Katz, Trans. ASSP-35, pp. 400-401, March 1987
11. F. Kubala, et al., ICASSP-88, pp. 291-294
12. W. Chou, et al., Proc. ICASSP-94, Vol. II, pp. 153-156