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

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

[†]NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo 180, JAPAN
[††]Tokyo Institute of Technology
[†††]Waseda University
matsuoka@splab.hil.ntt.co.jp

ABSTRACT

A large-vocabulary continuous-speech recognition (LVCSR) system was developed and evaluated. To evaluate the system, a Japanese business-newspaper speech corpus was designed and recorded. The corpus was designed so that is can be used for Japanese LVCSR research in the same way that the Wall Street Journal (WSJ) corpus, for example, is used for English LVCSR research. Since Japanese sentences are written without spaces between words, a morphological analysis was introduced to segment sentences into words so that word n-gram language models could be used. To enable the use of detailed word n-gram (n≥3) language models, a two-pass decoding strategy was applied. Context-dependent (CD) phone models and word trigram language models reduced the word error rate from 80.2% to 10.1% (an error reduction of about 88%). This result shows that CD phoneme modeling and word trigram language models can be used effectively in Japanese LVCSR.

1. INTRODUCTION

Large-vocabulary continuous speech recognition (LVCSR) is being extensively studied for the American and British English, French, German, and Italian languages using vocabularies taken from business newspapers such as the Wall Street Journal (WSJ) [1-9]. However, no similar research has been reported for the Japanese language. 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 useful for LVCSR. To enable word n-grams to be used for Japanese LVCSR, we have introduced a morphological analysis to segment Japanese sentences into words (morphemes).

To evaluate our recognition system, we designed a speech corpus that could be used for Japanese LVCSR research [10], since no Japanese speech-corpus comparable to the WSJ corpus existed. We designed a business-newspaper speech corpus using articles taken from about five years of the Nihon Keizai Shimbun (the Nikkei newspaper). A word frequency list of 623k words was derived from 6.8M sentences. Vocabularies of 7k, 30k, and 150k were defined, which provides the same coverage as the vocabulary sizes (5k, 20k, and 64k) in the WSJ corpus.

We studied acoustic modeling and language modeling for Japanese LVCSR using this speech corpus [10]. For 7k vocabulary, the word error rate was 82.8% when context-independent acoustic models and no language models were used. This improved to 10.1% when both context-dependent acoustic models and trigram language models were used. This result proved that an n-gram language model works well in Japanese LVCSR.

2. DESIGN OF THE CORPUS

Newspaper articles from five-year period were divided into two parts: a four-years-and-nine-months component for training and a three-month component for testing.

2.1 Text preprocessing

Texts were preprocessed before the morphological analysis. This was done to keep the sentences easy to read and to avoid morphological-analysis errors so that language models could be realistically estimated. Since our target is LVCSR and not sentence dictation, we discarded marks (e.g. quotation marks, bullets, etc.) that are usually not pronounced in spoken communication.

Sentences that were too long were discarded from the training and test sets, because long sentences are more difficult to read. We assumed that 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 deviation was 13.8 words for the training set and 13.5 words for the test set. Sentences with lengths in terms of the number of words that lay in the range of the mean $\pm 2\sigma$ were used to make a word frequency list and for training n-gram language models. The test set consisted of sentences whose lengths were in the range of the mean $\pm\sigma$ to ensure readability.

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

2.2 Morphological analysis

Table 1 Comparison of Nikkei and WSJ

	Nikkei (Japanese)	WSJ (English)
Training text size	180 M	37.2 M
Distinct words	623 k	165 k
5k coverage	88.0 %	90.6 %
7k coverage	90.3 %	-
20k coverage	96.2 %	97.5 %
30k coverage	97.5 %	-
65k coverage	99.0 %	99.6 %
150k coverage	99.6 %	-
20k OOV rate	3.8 %	2.5 %

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

The segmentation into words required sophisticated morphological analysis. Our morphological analyzer has a dictionary of 250k morphemes, and the accuracy of the morphological analysis was 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 that appeared in the training set. We had a list of 623k words. Since the morphological analyzer has a dictionary of 250k words, 373k out of the initial 623k words were analyzed and designated as unknown words. Most of these unknown words were proper nouns or unusual technical terms.

To eliminate morphological 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 then discarded.

2.3 Design of the corpus

Table 1 shows the coverage of our recognition task and the WSJ recognition task. There are more distinct words used in the Nikkei than in the WSJ. Compound-words, which are particularly common in Japanese, 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 the Nikkei task is higher than that for the WSJ task. We defined 7k and 30k vocabularies to have the same coverage as the 5k and 20k vocabularies for the WSJ task.

To evaluate an LVCSR system, we defined five subsets 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 these subsets. The OOV words were limited to those appearing in the 150k-word 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 theexperiments 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.

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	-	-	-	-	-	-

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 have been 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. We used 42 phoneme classes, including silence. In these experiments, continuous speech recognition using a simple phoneme-loop grammar network was carried out with

Table 4 Test-set perplexity

Vocabulary size	Language model	Test-set perplexity		
		Nikkei	WSJ	
			VP	NVP
7 k / 5 k	Unigram	482	-	-
	Bigram	49	80	118
	Trigram	27	44	68
30 k / 20k	Unigram	667	-	-
	Bigram	77	158	236
	Trigram	-	101	155

the phoneme defined 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. Here, for example Di2000 denotes a diphone-context-dependent model set whose training samples could be observed more than 2000 times in the training set, and Tri*Number* denotes a triphone-context-dependent model set. The highest accuracy of 61.6% was achieved when Di700 and Tri300 model sets were used together with context-independent (CI) models and smoothing was applied.

4. LANGUAGE MODELING

Word n-grams were estimated using the 6.8M sentences of the training set. Most of the bigrams and trigrams were singletons (they appeared only once in the entire training set). Since the average number of bigram and trigram occurrences were low, that is, we had only five to seven occurrences for each trigram on average, the language models obviously needed to be smoothed. We used the back-off smoothing method proposed by Katz [11].

Using the smoothed n-gram language models, we evaluated the test-set perplexity. The test-set sentences of the subsets 7k and 30k were used to calculate the test-set perplexity. Table 4 shows the test-set perplexity for the Nikkei task compared with that for the WSJ task [8]. The test-set perplexities were smaller for the Nikkei task than for the WSJ task.

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 for the WSJ task. We should note these different conditions.

5. TWO-PASS DECODING

To utilize word trigram language models, we applied a two-pass decoding approach. Since there were 2.1M bigrams and 17.1M trigram, it is inefficient to use all the n-gram models in a single search pass. Therefore, we used bigram models for the first pass and trigram models for the second pass. The first pass search produced N-best hypotheses using a simple word-loop grammar network and bigram models. Then, in the second-pass, the best hypothesis was found among the N-best hypotheses using trigram language models. The same acoustic models were used for the first and second pass search; that is, acoustic scores were kept after the first pass search and were re-used when necessary in the second pass along with the trigram language scores. The weighting factors for acoustic and language scoring were optimized experimentally for both the first pass and the second pass.

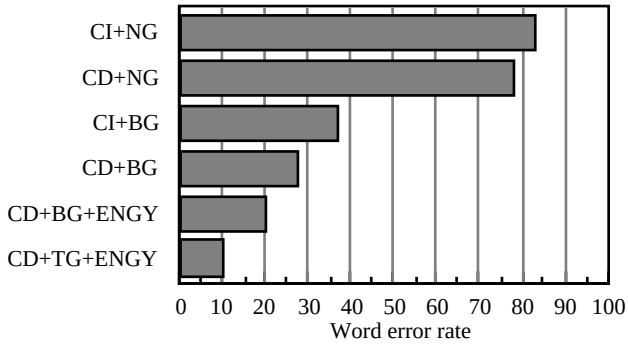

Figure 1 LV CSR experimantal results

6. 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. The context-dependent models consist of 502 triphone-context , 204 diphone-context, and 42 context-independent models. Figure 1 shows the LVCSR 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.

The word error rate for the baseline system with context-independent phoneme models and no language models (CI+NG) was 82.8% for the test set. This improved to 36.3% when the bigram language models were used (CI+BG). Using context-dependent phoneme models and introducing log-energy and Δ-log-energy further improved the word error rate to 20.0% (CD+BG+ENGY). Finally, word trigram language models were applied to find the best hypothesis among the N-best hypotheses produced by CD+BG+ENGY in the second-pass search, and this improved the word error rate to 10.1% (CD+TG+ENGY). In other words, the error rate was approximately halved when bigram language models were incorporated, and approximately halved again by the further addition of the context-dependent acoustic models with log-energy and Δ-log-energy. The word trigram again reduced the remaining error by approximately half.

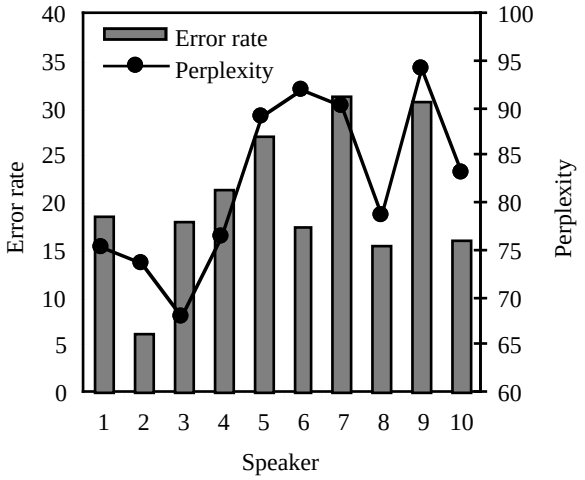

Figure 2 Word error rate and test-set perplexity for each speaker

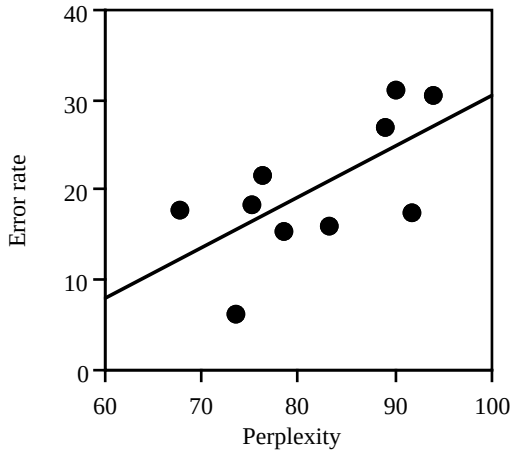

Figure 3 Relation between perplexity and word error rate

As the performance differed depending on the speakers, Figure 2 illustrates the results in the CD+BG+ENGY case for each speaker. Each speaker read different sentences, which varied in recognition difficulty, so the difference in performance could be attributed to this. 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 perplexity 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 derived from a Japanese business newspaper for LVCSR and have evaluated an LVCSR 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 LVCSR 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 language models, was 82.8%. This improved to 10.1% when context-dependent acoustic models and trigram language models were used. The error reduction, therefore, was 88%. Bigram and trigram language models and context-dependent acoustic models reduced the error rate very effectively. The bigram language models reduced the error by half. Similarly, the further addition of the context-dependent acoustic models again halved the remaining error. Finally, the trigram language models reduced the remaining error by another half.

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. And for the language models, we are introducing higher order n-gram models in addition to trigram and bigram models.

ACKNOWLEDGMENT

We are grateful to Nihon Keizai Shimbun Incorporated for allowing us to use the newspaper text database (Nikkei CD-ROM 90-94) for our research.

REFERENCES

1. D. B. Paul and M. Baker, "The design for the Wall Street Journal-based CSR corpus," Proc. ICSLP-92, pp. 899-902, Oct. 1992
2. J. L. Gauvain, L. F. Lamel, and M. Eskenazi, "Design considerations and text selection for BREF, a large French read-speech corpus," Proc. ICSLP-90, pp. 1097-1100, Oct. 1990
3. T. Robinson, J. Fransen, D. Pye, J Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," Proc. ICASSP-95, pp. 81-84, May 1995
4. H. J. M. Steeneken and D. A. van Leenwen, "Multi-lingual assessment of speaker independent large vocabulary speech-recognition systems: SQUALE-project," Proc. Eurospeech-95, pp. 1271-1274, Sep. 1995
5. P. C. Woodland, C. J. Leggetter, J. J. Odel, V. Valtchev, and S. J. Young, "The 1994 HTK large vocabulary speech recognition system," Proc. ICASSP-95, pp. 73-76, May 1995
6. D. Pye, P. C. Woodland, and S. J. Young, "Large vocabulary multilingual speech recognition using HTK," Proc. Eurospeech-95, pp. 181-184, Sep. 1995
7. L. Lamel. M. Adda-Decher, and J. L. Gauvain, "Issues in large vocabulary, multilingual speech recognition," Proc. Eurospeech-95, pp. 185-188, Sep. 1995
8. D. B. Paul and B. F. Necioglu, "The Lincoln large-vocabulary stack-decoder HMM CSR," Proc. ICASSP-93, pp. 660-663, Apr. 1993
9. L. Lamel and R. De Mori, "Speech recognition of European languages," Proc. IEEE Automatic Speech Recognition Workshop, pp. 51-54, Snowbird, Dec. 1995
10. T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui, and K. Shirai, "Large-vocabulary continuous-speech recognition using a Japanese business newspaper (Nikkei)," Proc. ARPA Speech Recognition Workshop, pp. 137-142, Feb. 1996
11. S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," Trans. ASSP-35, pp. 400-401, March 1987