論文 / 著書情報
Article / Book Information

Title	Large-Vocabulary Continuous-Speech Recognition Using a Japanese Business Newspaper (Nikkei)
Authors	Tatsuo Matsuoka, Katsutoshi Ohtsuki, Takeshi Mori, Sadaoki Furui, Katsuhiko Shirai
Citation	DARPA Speech Recognition Workshop Proceedings 1996 (Paperback), Vol. , No. , pp. 137-142
Pub. date	1996, 2

ARPA Speech Recognition Workshop

LARGE-VOCABULARY CONTINUOUS-SPEECH RECOGNITION USING A JAPANESE BUSINESS NEWSPAPER (*NIKKEI*)

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

Large-vocabulary continuous-speech recognition is being extensively researched for American/British English, French, German and Italian languages by using business newspapers. No research efforts, however, have been reported using Japanese newspapers. This is mainly because Japanese sentences are written without spaces between words. Therefore, it is difficult to use a word N-gram language model, which is very useful for large-vocabulary continuous-speech recognition. We studied Japanese large-vocabulary continuous-speech recognition using a Japanese business newspaper. To use word N-gram, 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. Each of 54 speakers uttered 100 sentences. Using the first 10 speakers' speech from 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 the bigram language models and the 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 using business newspapers such as the Wall Street Journal, Le Monde, Frankfurter Rundschau, and Sole 24 Ore [1-9]. No research efforts, however, have been reported for Japanese. This is mainly because Japanese sentences are written without spaces between words, and they are very difficult to automatically segment. Therefore, it is difficult to use a word N-gram language model, which is very helpful for LV CSR. There has been no speech corpus for Japanese LV CSR using business newspapers. Therefore, we first designed a speech corpus which can 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 600k 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. Each of 54 speakers uttered 100 sentences to build a speech corpus.

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 effect of context-dependent acoustic modeling and word N-gram language modeling.

2. DESIGN OF THE CORPUS

2.1 Text database

Newspaper articles for five years were divided into two parts: four years and nine months for training and three months for testing. There were 6.8M sentences and 180M words in the training set, and 342k sentences and 9.8M words in the test set after the text preprocessing described in the following section (Table 1).

2.2 Text preprocessing

Texts were preprocessed before morphological analysis. This was to keep the sentences easy to read and also to keep morphological analysis correct so that language models could easily be estimated. Our target is LV CSR and not dictating sentences, so we discarded marks which are usually not pronounced in spoken communications. Marks used for emphasis such as •, !, ? were omitted because their pronunciations are not well defined in Japanese, and are therefore difficult to read consistently. Quotation marks such as " ", ' ' were omitted, but the contents were kept because the contents are part of the sentence. Parentheses such as ( ), [ ], < > were omitted with the contents because they are used to give readings of difficult Kanji characters or meanings of difficult words, and the sentence structure does not change without them.

Sentences which were too long were discarded from the training and test sets. This is because long sentences are not

Table 1 Number of sentences and words in the text database

	Training set	Testing set
Number of sentences	7.3M / 6.8M	374K / 342K
Number of words	200M / 180M	10.7M / 9.8M

before/after the text preprocessing

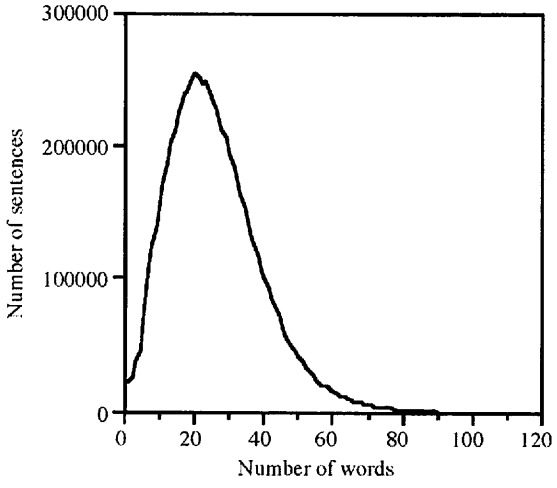

Figure 1 Distribution of sentence length in terms of the number of words in a sentence

easy to read. We assumed the distribution of the sentence length in terms of the number of words in a normal distribution. Figure 1 shows the actual distribution for the entire text database. The number of words in a sentence was 25.6 words 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±2σ in terms of the number of words in a sentence were used for making 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±σ to ensure readability.

2.3 Morphological analysis

Since there is no spacing between words in Japanese, words are not clearly delimited. To calculate statistics such as word frequency, we first segmented sentences into words. This segmentation requires sophisticated morphological analysis. Our morphological analyzer has a dictionary of 250K morphemes. The accuracy of the morphological analysis is about 95% for the newspaper articles that we used. In this study, we defined words by morphemes according to the lexicon of our morphological analyzer.

Table 3 Entry items corresponding to the number of homophone classes with *k* graphemic forms in the class

Corpus	Rate in Lexicon	Homophone class size (*k*)			
		1	2	3	≥4
Nikkei (30K)	20 %	24.1 K	2438	716	565
BREF (10 K)	35 %	6686	1329	215	73
BREF (40 K)	45 %	23.7 K	5361	1264	1039
WSJ (9 K)	6 %	8453	237	22	1
WSJ (65 K)	15 %	60.4 K	3689	890	291
FR (64 K)	10 %	58.1 K	2769	221	57
So24 (10 K)	1.7 %	9872	85	3	0

To define the vocabulary size, we first made a word frequency list. This is a frequency-based-sorted list of words appearing in the training set. As a result, we had a list of roughly 600K words. Since the morphological analyzer has a dictionary of 250K words, 350K out of 600K words were analyzed as unknown words. Most of the unknown words were proper nouns or very special technical terms.

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

2.4 Design of the corpus

We investigated the relation between vocabulary size and coverage for various languages. Table 2 shows the coverage of our task and the LV CSR tasks for other languages [9]. There are more distinct words for Japanese and German than for the other three languages. German is known as a language which has many compound-words. These compound-words result in a large number of distinct words. This is also true for Japanese. In addition, inflection increases the nominal number of distinct words in Japanese. The coverages for 5K and 20K vocabularies for Japanese are not so low, while the number of distinct words is much larger. The out-of-vocabulary rate for Japanese is higher than that for English but lower than that for German.

Table 2 Comparison of lexica and LM training corpora for different languages

	Nikkei (Japanese)	WSJ (English)	Le Monde (French)	Frankfurter Randchau (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 %
20K coverage	96.2 %	97.5 %	94.7 %	90.0 %	96.3 %
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 4 Vocabulary size and coverage

Vocabulary size	Coverage (%)	
	Nikkei	WSJ
5K	88.0	91.7
7K	90.3	-
20K	96.2	97.7
30K	97.5	-
64K	98.9	99.6
150K	99.6	-
173K	99.7	100.0
623K	100.0	-

Table 5 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

Table 6 Percentage of complete sentences made by each subset

Subset	Training set	Testing set
7K	10.7	11.1
7K, 7K+	40.8	41.8
30K	42.0	42.9
30K, 30K+	62.0	62.2
30K, 30K+, 30K++	64.2	63.9

Table 7 Average number of words and duration for each sentence

Subset	Number of words	Duration [s]
7K	20.9	6.3
7K+	23.3	7.0
30K	23.4	7.1
30K+	25.7	7.9
30K++	27.4	8.6

Table 3 lists the number of homophones in each corpus [9]. The homophone-class size k=1 means these words do not have homophones. The homophone-class size k=2 means that the word has one homophone. According to Table 3, Japanese and French have more homophones than other languages. When the homophone rate in the lexicon is high, it is more difficult to perform LV CSR for that language, because it is difficult to extract the meanings without a semantic knowledge source.

Table 8 Phonemes

Vowels	a, e, i, o, u aa, ee, ei, ii, oo, ou,uu
Consonants	b, d, g, p, t, k ch, j, sh, ts, f, h, s, z N, m, n, r, w, y by, gy, hy, ky, my, ny, py, ry
Double consonant	Q
Silence	#

Table 4 shows a detailed comparison of the coverage for the Nikkei task and 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 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 test sets. The description of subsets is listed in Table 5. The OOV words are limited to those appearing in the top 150K vocabulary, that is, our task is limited to 150K words.

Table 6 lists the percentage of sentences in the training and test sets that can be completed by the vocabulary allowed in each subset. Although the subset 30K++ allows more than two OOV words in a sentence, only 64% of the sentences can be completed using the vocabulary.

Each of 54 speakers uttered 100 sentences, 20 sentences from each of the subsets. Fifty sentences were selected form the training set, and fifty sentences were selected from the test set. Speech was recorded through a head-mounted Sennheiser (HMD-410) microphone and a desk-mounted Crown (PCC-160 phase coherent cardioid) microphone simultaneously. Speech recorded through the head-mounted microphone was used for the experiments described in this paper. Table 7 lists the average number of words in a sentence and the average time duration of a sentence for each subset. As the number of vocabulary words increases the length increases, although not significantly.

3. ACOUSTIC MODELING

It is reasonable to use subword (for example, phoneme) acoustic models for LV CSR, because we usually cannot afford word-based models for a large vocabulary. Table 8 lists the Japanese phonemes we used. There were 42 phonemes including silence.

We evaluated context-independent, 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

Table 9 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 10 Number of N-grams in the training set

	Unigram	Bigram	Trigram
7 K	7000	2.1 M	17.1 M
30 K	29991	4.9 M	30.5 M

Table 11 Average number of occurrences of each N-gram

	Unigram	Bigram	Trigram
7 K	24388.0	65.1	7.2
30 K	6121.9	33.8	5.1

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. For training the initial acoustic models, we used phonetically-balanced-sentence speech which is phoneme labeled. Then, we performed embedded training using different phonetically-balanced-sentence speech and speech from an information service task. In total, more than 15,000 utterances from 58 speakers were used to train the acoustic models.

Each phone model except the silence model has three states and three loops. The silence model has one state and one loop. Each state has four mixture Gaussian distributions. Speech is sampled at 12kHz and digitized into 16 bits. The frame length is 32 ms and the shift is 8 ms. The acoustic features used are 16 LPC derived cepstra and log-energy, and their first derivatives (delta-features).

To cope with co-articulation problems, context-dependent models were used. Diphone-context-dependent models consider left context. The context-dependent models were trained using the context-independent models as seed models. Table 9 lists the results of the phoneme recognition experiments. 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 can be observed more than 2000 times in the training set. Tri600 means triphone-context-dependent models whose training samples can be observed more than 600 times in the training set. When Di700 and Tri300 were used with context-independent (CI) models with smoothing, the highest accuracy of 61.6% was achieved.

4. LANGUAGE MODELING

When the vocabulary size is large, a language model is indispensable. There are three major types of language models: finite state network, context free grammar (CFG) and N-gram. For LV CSR, word N-grams are usually used, because when the vocabulary size is large it is difficult to build a compact and encompassing finite state network or to write a good CFG manually. On the other hand, word N-grams can easily be estimated if a large amount of text is available.

Table 10 lists the number of unigrams, bigrams, and trigrams observed in the training set. Table 11 lists the average number of occurrences of each corresponding N-gram. Most of the bigrams and trigrams are singleton (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, the language models must obviously be smoothed. We applied the back-off

Table 12 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

Table 13 Large-vocabulary continuous speech recognition results

Acoustic model		Language model	Training set		Test set	
HMM	Features		% Correct	Accuracy	% Correct	Accuracy
CI	Cepstrum Δcepstrum	NG	19.6	18.0	18.1	17.2
CD		NG	24.5	23.0	23.1	22.1
CI		BG	67.0	65.0	65.8	63.7
CD		BG	77.2	73.5	76.0	72.5
CI	+ log-energy Δlog-energy	BG	75.3	73.4	73.2	71.5
CD		BG	82.4	80.3	82.3	80.0

CI: context-independent, CD: context-dependent, NG: no-grammar, BG: bigram

Table 14 Results for each speaker (Test set)

Speaker	Test-set perplexity	CD + NG		CD + BG		CD + log-energy + BG	
		% Correct	Accuracy	% Correct	Accuracy	% Correct	Accuracy
m001	75.3	28.9	28.9	85.3	79.6	84.4	81.5
m002	73.5	30.9	30.9	92.1	91.6	95.0	93.8
m003	67.8	23.5	23.5	76.5	72.6	83.8	82.1
m004	76.3	25.8	24.7	80.1	79.0	81.2	78.5
m005	88.9	20.6	20.6	75.9	74.7	75.9	72.9
m006	91.7	16.8	15.3	71.1	67.9	82.6	82.6
m007	90.2	13.6	11.6	58.3	49.8	72.9	68.8
m008	78.5	25.7	23.7	79.5	75.9	85.6	84.6
m009	94.0	18.8	17.2	67.7	62.4	74.7	69.4
m010	83.1	25.2	24.3	73.4	72.0	84.4	83.9
Average	81.8	23.1	22.1	76.0	72.5	82.3	80.0

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 (articles from the last three months were assigned as the test set) were used to calculate test-set perplexity. Table 12 shows the test-set perplexity for the Nikkei task and 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 are different for our task and the WSJ task. Regardless of these differences in details, the perplexities are relatively comparable.

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 for the context-dependent models, a 748-model set was used because this achieved the best result in the preliminary phoneme recognition experiments. Table 13 lists the LV CSR results. The percentage correct and accuracy were obtained as

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

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

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

As shown in Table 13, the accuracy for the baseline system with context-independent phoneme models and no-grammar language models was 17.2% for the test set. This improved to 63.7% by using the bigram language models. By using context-dependent phoneme models and introducing log-energy and Δ-log-energy, it improved to 80.0%. The error rate was reduced by approximately half with the incorporation of the bigram language models. 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, Table 14 lists the results for each speaker. Each speaker read different sentences, and the difficulty of the sentences to be recognized

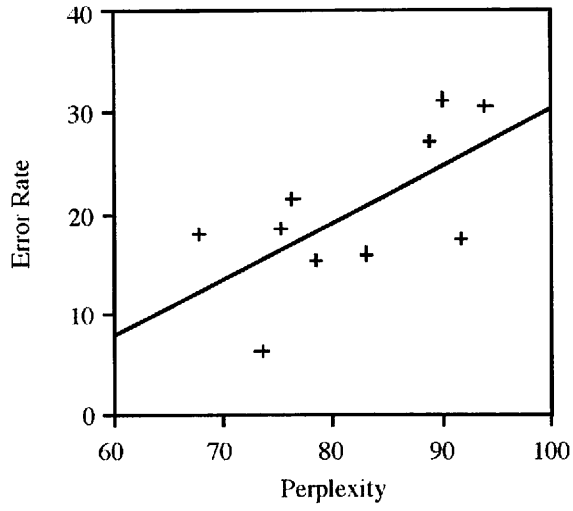

Figure 2 Relation between perplexity and word error rate

varied, 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 Table 14, the test-set perplexity varied from 67.8 to 94.0. Figure 2 illustrates the relation between the test-set perplexity and the word error rate. The results show that the error rate increases as the perplexities increase. The solid line indicates the first order regression. The deviation from the line can be interpreted as variability due to acoustical characteristics.

6. SUMMARY

We have described the design of a speech corpus using 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% using context-dependent acoustic models and bigram language models. The error reduction, therefore, is 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. As for the acoustic models, inter-word context-dependent models are currently being introduced instead of intra-word context-dependent models [12]. As for the language models, we are introducing trigram models in addition to bigram and unigram models.

ACKNOWLEDGMENT

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

REFERENCES

1. D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR corpus," Proc. ICSLP-92, pp. 899-902
2. J. 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
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
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
5. P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev, and S. J. Young, "The 1994 HTK Large Vocabulary Speech Recognition System," Proc. ICASSP-95, pp. 73-76
6. D. Pye, P. C. Woodland, and S. J. Young, "Large Vocabulary Multilingual Speech Recognition using HTK," Proc. EUROSPEECH-95, pp. 181-184
7. L. Lamel, M. Adda-Decher, and J. L. Gauvain, "Issues in Large Vocabulary, Multilingual Speech Recognition," Proc. EUROSPEECH-95, pp. 185-188
8. D. B. Paul, and B. F. Necioglu, "The Lincoln Large-Vocabulary Stack-Decoder HMM CSR," Proc. ICASSP-93, pp. 660-663
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. 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
11. F. Kubala, et al., "Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000-Word Resource Management Database," ICASSP-88, pp. 291-294
12. W. Chou, T. Matsuoka, B.-H. Juang, and C.-H. Lee, "An Algorithm of High Resolution and Efficient Multiple String Hypothesization for Continuous Speech Recognition Using Inter-Word Models," Proc. ICASSP-94, Vol. II, pp. 153-156