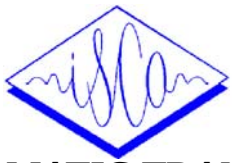
ISCA Archive
http://www.isca-speech.org/archive



EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

TOWARD AUTOMATIC TRANSCRIPTION OF JAPANESE BROADCAST NEWS

Tatsuo Matsuoka†, Yuichi Taguchi††, Katsutoshi Ohtsuki†, Sadaoki Furui†††, and Katsuhiko Shirai††
†NTT Human Interface Laboratories
1-1 Hikari-no-Oka, Yokosuka-shi, Kanagawa 239, Japan
††Waseda University, †††Tokyo Institute of Technology

10.21437/Eurospeech.1997-325

ABSTRACT

In this paper, we report on the automatic recognition of Japanese broadcast-news speech. We have been working on large-vocabulary continuous speech recognition (LVCSR) for Japanese newspaper speech transcription and have achieved good performance. We have recently applied our LVCSR system to transcribing Japanese broadcast-news speech. We extended the vocabulary from 7k words to 20k words and trained the language models using newspaper texts and broadcast-news manuscripts. These two language models were applied to our evaluation speech sets. The language model trained using broadcast-news manuscripts achieved better results for broadcast-news speech than the language model trained using newspaper texts, which achieved better results for newspaper speech. We achieved a word error rate of 19.7% for anchor-speaker's speech by using a bigram language model and a trigram language model both trained using broadcast-news manuscripts.

1. INTRODUCTION

The DARPA Hub-4 test that began in 1995 is evaluating the use of LVCSR (large-vocabulary continuous speech recognition) to transcribe audio recordings of broadcast news. Several preliminary Hub-4 evaluation results have been reported [1-5]. Coincidentally, in 1996, the Japanese government announced that it will issue a regulation in several years requiring TV news programs to be closed captioned. Transcribing broadcast news is a challenging task, and thus a good test of applying LVCSR technology to real-world systems. We are therefore investigating the automatic recognition of Japanese broadcast-news speech. This paper describes some of our preliminary results.

We have been working on LVCSR for read newspaper speech. So far, a word error rate of about 10% has been achieved for a 7k-word vocabulary [6-8]. Figure 1 shows the progress of our LVCSR performance for newspaper speech recognition. We found that bigram and trigram language models are very effective for Japanese LVCSR. Our trigram language model reduced the word error rate from 18.1 % to 10.1%. This improvement is much larger than those for other languages. We also had better results with acoustic models designed using tree-based clustering. A word error rate of 9.5% was obtained with those acoustic models and trigram language models.

We have applied our LVCSR system to transcribing Japanese broadcast-news speech. We extended the vocabulary to 20k words and trained the language models by using newspaper texts and broadcast-news manuscripts. We conducted phoneme-recognition experiments to examine if broadcast-news speech is acoustically more difficult than read newspaper speech. We then experimentally compared two language models: one trained using broadcast-news manuscripts and one trained using newspaper texts.

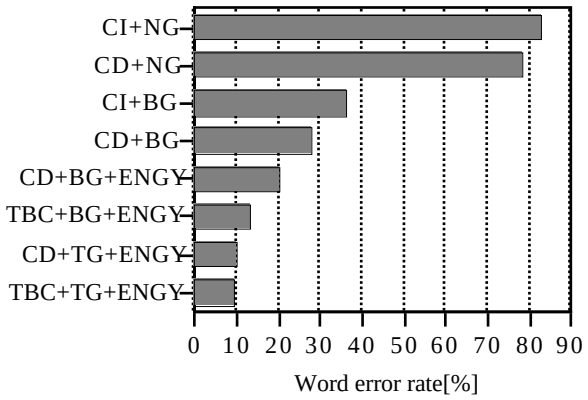
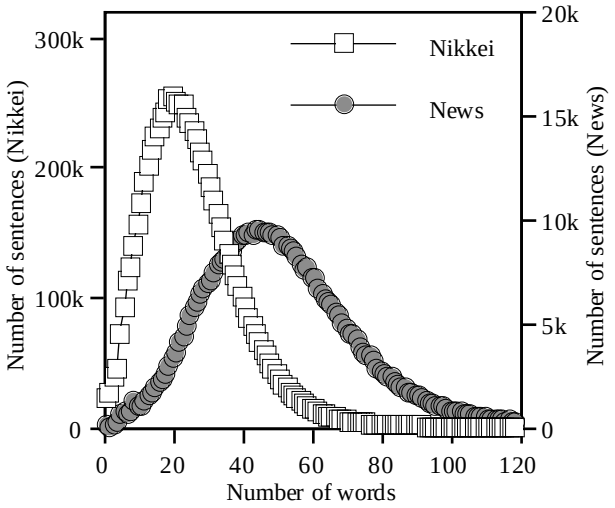


Figure 1: LVCSR experimental results for newspaper speech

CI: context-independent acoustic models,
CD: context-dependent acoustic models,
NG: no grammar models,
BG: bigram language models,
TG: trigram language models,
ENGY: energy parameters added to the feature parameters,
TBC: tree-based clustering used in designing acoustic models.

2. BROADCAST NEWS DATA

Raw audio recordings of broadcast news include frequent speaker changes, background music, and telephone speech, such as field reports. We segmented these parts manually and used only the clean-speech parts, i.e., those parts not containing background music, noises, telephone speech, or overlapped speech for the experiments reported here. These experiments correspond to the partitioned evaluation (PE) with the baseline broadcast (F0) condition in the 1996 Hub-4 test [12]. Even using only clean speech is still challenging because news speech is usually much more fluent than read speech and includes spontaneous speech



Figire 2: Histgram of the number of words per sentence

phenomena, such as '*uh*' at the beginning of a sentence or the correction of slips. Furthermore, we found that the sentences are much longer for broadcast news than for newspapers. Figure 2 shows distributions of the number of words per sentence in broadcast-news manuscripts and newspaper texts. The average number for the broadcast-news is about 50 words, double that for the newspaper.

To apply n-gram language models, we segmented the broadcast-news manuscripts into words by using a morphological analyzer because Japanese sentences are written without spaces between words. Some of the irrelevant symbol-marks, such as brackets were filtered out. The manuscripts have typographical errors and unread comments such as '*with VTR*' and some of them could be corrected or removed automatically. A word-frequency list was derived from the filtered sentences, and the 20k most frequently used words were selected as the vocabulary words. This 20k vocabulary covers about 98% of the words in the broadcast-news manuscripts. Table 1 lists the training-text size and the coverage for broadcast-news, the Nikkei newspaper and the Wall Street Journal. The broadcast-news manuscripts were from August 1992 to May 1996 and the newspaper texts were from January 1990 to September 1994. Although both training texts extended over a period of about 5 years, the training-text size for the broadcast-news is noticeably smaller than for the newspapers.

Table 1: Comparison of lexica and LM training

	News	Nikkei	WSJ
Training text size (words)	24M	180M	237M
Number of distinct words	114k	623k	476k
5k coverage	91.5%	88.0%	90.6%
7k coverage	93.7%	90.3%	-
20k coverage	98.0%	96.2%	97.5%
30k coverage	99.1%	97.5%	-
65k coverage	99.7%	99.0%	99.6%

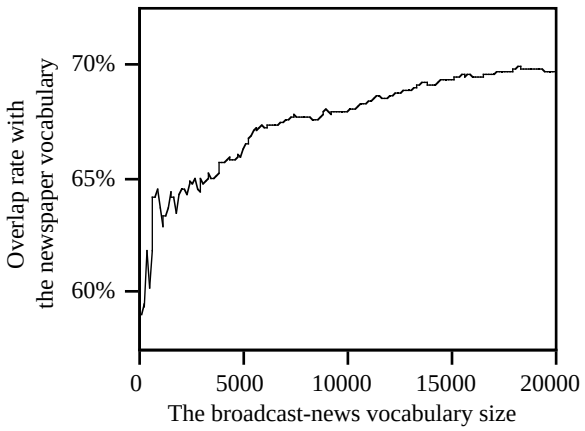


Figure 3: Overlapping rate of the broadcast-news vocabulary and the newspaper vocabulary

Figure 3 shows the overlap rate of the broadcast-news vocabulary and the newspaper vocabulary, which have the same vocabulary size. The broadcast-news 20k vocabulary words and the newspaper 20k vocabulary words have about a 70% overlap, or about 14k words of the each 20k vocabulary are identical.

3. ACOUSTIC MODELING

The acoustic models we used were all shared-state context-dependent phoneme HMMs designed using tree-based clustering [9]. The total number of states was 2,106, and the number of Gaussian mixture components per state was 4. They were trained using phonetically-balanced sentences and dialogue read by 53 speakers. All of the speakers were male, thus the HMMs were gender-dependent models. The total number of utterances was 13,270 and the total volume of training data was approximately 20 hours.

To investigate the acoustical difference between broadcast-news speech and read newspaper speech, we conducted phoneme-recognition experiments. Table 2 shows the results. The percent correct and the accuracy were calculated as follows.

$$\%Correct = \frac{N - sub. - del.}{N} \cdot 100$$

$$Accuracy = \frac{N - sub. - del. - ins.}{N} \cdot 100$$

We also investigated the speech rate of speech data by counting the number of phonemes per second. The broadcast-news speech had 12.4 phonemes per second and the read newspaper speech had 12.6. It was found that the two types of speech data had almost the same speech rates. The accuracy of phoneme recognition was almost same for the broadcast-news speech and the read newspaper speech.

4. LANGUAGE MODELING

As shown in Figure 1, n-gram language models have proven to be very effective in Japanese LVCSR for read newspaper speech

[6-8]. We can expect the same effectiveness for news-speech transcription. To train n-gram language models, we need a large amount of text data. As Table 1 shows, collecting a large amount of data is usually easier for newspaper text than it is for broadcast-news manuscripts. Therefore, it would be helpful if a newspaper language model also worked well for broadcast news.

To determine if a newspaper language model can be used for broadcast news, we trained two language models, one using broadcast-news manuscripts and one using newspaper texts. Table 3 shows the distinct number and the average occurrence of word n-gram models in each training text. Due to the small training-text size, the distinct number of bigrams for broadcast news was a quarter of that for the Nikkei newspaper and smaller than one fifth for trigrams. We estimated unseen n-gram probabilities using Katz's back-off smoothing method [10]. More n-grams had the estimated value for broadcast-news language models than for the newspaper models.

5. LVCSR EXPERIMENTS

The evaluation speech sets are summarized in Table 4. We divided the news-speech data set, which was broadcasted on TV in June 1996, into two parts: one for anchor speakers and one for other speakers. For a comparison, we also used a read-newspaper-speech set that had a 30k-vocabulary [6-8].

The LVCSR results with bigram language models are shown along with the test-set perplexities in Table 5. The word error rate was calculated as follows.

$$WordErrorRate = \frac{sub. + del. + ins.}{N} \cdot 100$$
$$= 100 - Accuracy$$

Each of the two language models was applied to each evaluation speech set. The News LM achieved better results for news speech (Anchor and Others) than the Nikkei LM, which achieved better results for read newspaper speech (Nikkei). To investigate why the News LM showed poor performance for Others, we plotted word error rate for each speaker against test-set perplexity (Figure 4 and 5). We found that the word error rate depends on test-set perplexity, not the type of speaker (See Fig. 5). The Others test set includes weather and sports news, thus the test set has high OOV rate and high perplexity.

We applied trigram language models to the anchor speakers' speech transcription. We used multi-pass search strategy [11] to apply higher order language models such as trigram models with less computational cost. In the first pass, N-best hypotheses for an utterance are computed using bigram language models. Those hypotheses are rescored using trigram language models and the most likely alternative was chosen as the recognition result. In the experiments, 300-best hypotheses were generated in the first pass and the same acoustic models were used for both of the passes.

Table 2: Phoneme recognition results for news speech and read newspaper speech

	News	Nikkei
%Correct	82.0%	80.7%
Accuracy	61.5%	64.7%

Table 3: Number and average occurrence of distinct n-grams

	n-gram	Distinct no.	Av. occurrence
Broadcast news	unigram	20k	1160
	bigram	0.9M	24
	trigram	4.4M	5
Nikkei newspaper	unigram	20k	8747
	bigram	3.6M	44
	trigram	24M	6

Table 4: Evaluation speech data

	Anchor	Others	Nikkei
No. of speakers	5	6	10
No. of utterances	99	121	100
No. of words	4159	2275	2168
20k-OOV rate	1.5%	3.7%	3.5%

Table 5: LVCSR results with bigram LM

Task	Language model	Test-set perplexity	Word error rate
News (Anchor)	News LM	105	23.7%
	Nikkei LM	190	31.5%
News (Others)	News LM	255	38.2%
	Nikkei LM	281	40.2%
Nikkei (30k)	News LM	253	31.0%
	Nikkei LM	100	22.8%

Table 6: LVCSR results with tirgram LM (Anchor)

Language model	Test-set perplexity	Word error rate	Error reduction from bigram
News LM	48	19.7%	16.9%
Nikkei LM	63	20.7%	12.7%

Table 6 shows the LVCSR results of the broadcast-news speech (Anchor). The first pass was computed with the News bigram language models and the 300-best hypotheses were rescored with either of the two trigram language models. In spite of a small amount of training text, the broadcast-news trigram language models, like the bigram news models, achieved a better result. The broadcast-news manuscripts were taken from not only the same task domain but also from a period of time closer to the test data.

6. CONCLUSION

In preliminary experiments on Japanese broadcast-news transcription, we have achieved a word error rate of 23.7% (bigram) and 19.7% (trigram) by using n-gram language models trained

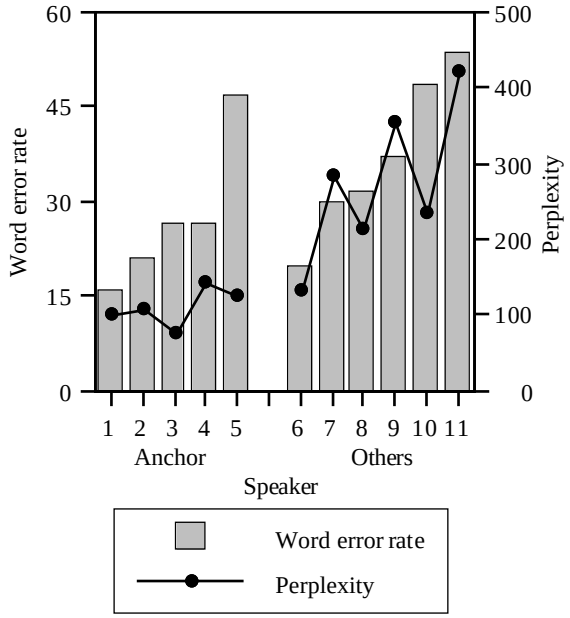


Figure 4: LVCSR results

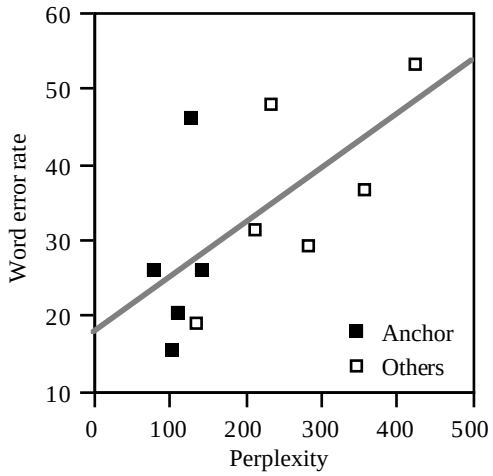


Figure 5: Perplexity vs. Word error rate

using broadcast-news manuscripts for anchor-speakers' speech. Phoneme-recognition experiments showed that acoustic models trained using read speech had almost the same performance for read newspaper speech and the broadcast-news speech. We can expect better results with acoustic models trained using broadcast-news speech. A newspaper language model was not as effective as a broadcast-news language model for broadcast-news transcription. It was found that a certain amount of training data from the same task domain and from a period of time closer to the test data provided a better language model than a very large amount of training data from another task. To use the newspaper language model effectively, a language model interpolation or adaptation method is definitely needed. The high OOV rate of Others may make the word error rate high, thus the unknown word problem must be dealt with . Furthermore Figure 5 shows that even in small ranges of perplexity, the range of the word error rate for the speakers is large. Speaker adaptation will improve performance for speakers with poor results.

References

[1] F. Kubala, T. Anastasakos, H. Jin, J. Makhoul, L. Nguyen, R. Schwartz, and N. Yuan, "Toward Automatic Recognition of Broadcast News," Proc. DARPA Speech Recognition Workshop, pp. 55-60, February 1996.

[2] U. Jain, M. A. Siegler, S. J. Doh, E. Gouvea, J. Huerta, P. J. Moreno, B. Raj, and R. M. Stern, "Recognition of Continuous Broadcast News with Multiple Unknown Speakers and Environments," Proc. DARPA Speech Recognition Workshop, pp. 61-66, February 1996.

[3] S. Wegmann, L. Gillick, J. Orloff, B. Peskin, R. Roth, P. van Mulbregt, and D. Wald, "Marketplace Recognition using Dragon's Continuous Speech Recognition System," Proc. DARPA Speech Recognition Workshop, pp. 67-71, February 1996.

[4] P. S. Gopalakrishnan, R. Gopinath, S. Maes, M. Padmanabhan, L. Polymenakos, H. Printz, and M. Franz, "Transcription of Radio Broadcast News with the IBM Large Vocabulary Speech Recognition System," Proc. DARPA Speech Recognition Workshop, pp. 72-76, February 1996.

[5] F. Kubala, T. Anastasakos, H. Jin, L. Nguyen, and R. Schwartz, "Transcribing Radio News," Proc. ICSLP-96, pp. 598-601, October 1996.

[6] T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui, and K. Shirai, "Large-Vocabulary Continuous Speech Recognition using a Japanese Business Newspaper (Nikkei)," DARPA Speech Recognition Workshop, pp. 137-142, February 1996.

[7] T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui, and K. Shirai, "Japanese Large-Vocabulary Continuous Speech Recognition using a Business-Newspaper Corpus," ICSLP-96, pp. 22-25, October 1996.

[8] T. Matsuoka, K. Ohtsuki, T. Mori, K. Yoshida, S. Furui, and K. Shirai, "Japanese Large-Vocabulary Continuous Speech Recognition using a Business-Newspaper Corpus," ICASSP-97, pp. 1803-1806, April 1997.

[9] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based State Tying for High Accuracy Acoustic Modeling," Proc. ARPA Human Language Technology Workshop, pp. 307-312, March 1994.

[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] R. Schwartz, L. Nguyen, and J. Makhoul,"Multiple-Pass Search Strategies," in Automatic Speech and Speaker Recognition, ed. C.-H. Lee et al., pp.429-456, 1996.

[12] J. S. Garofolo, J. G. Fiscus, and W. M. Fisher, "Design and Preparaton of the 1996 Hub-4 Broadcast News Benchmark Test Corpora," Proc. DARPA Speech Recognition Workshop, February 1997.