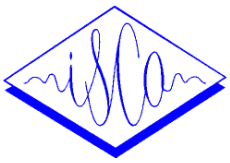

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

EFFECTS OF WORD STRING LANGUAGE MODELS ON NOISY BROADCAST NEWS SPEECH RECOGNITION

Kazuyuki TAKAGI Rei OGURO Kazuhiko OZEKI

The University of Electro-Communications, Chofu, Tokyo, 182-8585 Japan

http://www-oz.cs.uec.ac.jp/

ABSTRACT

In this paper, we present the results that our n-gram based word string language model, combined with speaker and noise adaptation of the acoustic model, improves recognition performance of noisy broadcast news speech. The focus was brought into a remedy against recognition errors of short words. The word string language models based on POS and n-gram frequency reduced deletion errors by 17%, insertion errors by 20%, and substitution errors by 3% in Japanese TV broadcast news speech recognition.

1 Introduction

Recently large vocabulary continuous speech recognition systems have been challenged with increasingly difficult tasks, as the research focus has shifted from read speech data to speech data found in the real world. The Japanese counterpart of HUB-4 [1] broadcast news recognition projects have been pursued by NHK (Japan Broadcasting Corporation) and other research institutes since 1996 [2, 3]. The research has now stepped in the stage where real time captioning of news speech [4] is being tested in real broadcasts, although the system limits its input to utterances of main announcers in studio clean environments.

There can be an approach to noisy broadcast news speech recognition from the language modeling side as well as the acoustic side. To get a language model that is robust against noise is as important as to get a language model that provides low test-set perplexity for realizing a good broadcast speech recognition system. Many Japanese LVCSR systems adopt morphemes as the units for recognition among which there are key morphemes such as particles, copulas and auxiliary verbs that are very short consisting of only one or two syllables. This suggests that those short morphemes are easy to be deleted or substituted in the recognition process, although they play key grammatical roles in a sentence.

The following work shows the result that our word string language models based on part of speech and frequency criteria alleviate the problem and improve recognition performance for noisy broadcast news speech.

2 Baseline

2.1 Broadcast news corpus

A Japanese broadcast news corpus was provided by NHK [2], which contains the manuscripts and the broadcast audio recordings of two morning news programs, one noon news program, and one evening news program. The manuscript corpus is a collection of broadcast news articles that spans over 52 months between August 1992 and August 1997, among which the first 50 months (837K sentences, 41.8M words) were used for language model training, and the manuscript of July 1996 (23K sentences, 3.9M characters) were held out for evaluation of the language models.

The evaluation speech data set of 160 sentences was selected from the audio recordings from July 1 to July 14, 1996. The evaluation set contains two subsets of 80 male announcer sentences and 80 female announcer sentences, each of which has the equal number of utterances in studio clean and in noisy environments. Those sentences were so selected that they have no OOV with reference to the baseline language model in order to avoid the OOV issues on speech recognition results.

2.2 Baseline language model

The manuscript texts of the first 50 months in corpus were first split into morphemes (hereafter, "words") by Japanese morphological analyzer ChaSen [8]. Each word is presented as a pair of its orthographic representation and POS tag. According to a word frequency list, 18.8k most frequent words in training corpus were selected as the baseline vocabulary (cutoff frequency = 34; word coverage = 98%). Then the trigram language model was generated using CMU-Cambridge SLM [9], with Good-Turing backoff smoothing method (cutoff = 1). This is the baseline language model.

3 Word string language model

Most Japanese LVCSR systems use morpheme as a recognition unit. However it is questionable whether a morpheme is the best basic unit for stochastic language modeling. Several reports indicated that grouping frequent word strings into new words improves the language model performance [5, 6, 7].

3.1 Log likelihood ratio based selection

This is based on the log likelihood ratio (LLR) of two hypotheses on the bigram (w_1, w_2): H_1 (occurrence of w_2 is independent of the previous occurrence of w_1) and H_2 (occurrence of w_2 is dependent on the previous occurrence of w_1).

$$\begin{aligned} LLR(w_1, w_2) &= \log \frac{L(H_1)}{L(H_2)} \\ &= (\sum_{ij} c_{ij}) \log(\sum_{ij} c_{ij}) + \sum_{ij} (c_{ij} \log c_{ij}) \\ &\quad - \sum_{i} ((\sum_{j} c_{ij}) \log(\sum_{j} c_{ij})) \\ &\quad - \sum_{j} ((\sum_{i} c_{ij}) \log(\sum_{i} c_{ij})) \end{aligned}$$

where $c_{11} = F(w_1, w_2)$, $c_{12} = F(w_1) - c_{11}$, $c_{21} = F(w_2) - c_{11}, c_{22} = N - c_{11} - c_{12} - c_{21}$ (N: total number of word tokens in the corpus).

It is reported in the literature that for the read speech material of JNAS (Japanese News Article Sentences) the word string language model created by this measure exhibits best performance both in test set perplexity and speech recognition evaluation, compared with other measures, i.e. co-occurrence frequency, mutual information and entropy [10].

The same measure was tested on the broadcast news data. Log likelihood ratio of the bigrams was calculated whose occurrence frequency exceeds the baseline cutoff threshold. Then word string sets of three different sizes were created by taking the top 500, 1080, and 2160 bigrams from the list. The word string language models, LLR500, LLR1080, and LLR2160 were trained on the manuscript corpus, where each token of the word strings was concatenated into a single word. The vocabulary for a word string language model was created by adding the selected word strings to the baseline lexicon. The word string model based on log likelihood ratio showed lower test set perplexity, as anticipated, than the other word string models as in Table 2, with the lowest test set perplexity given by LLR1080.

3.2 POS and frequency based selection

3.2.1 Function word string (FNC)

Looking at the news manuscript corpus, we noticed that certain types of morpheme strings occur very frequently, and that most of them contain very short morphemes that consist of only one or two syllables, which tends to invoke errors especially in noisy speech. Among those are function word strings. For instance, formal and politeness expressions that are used commonly in broadcast news speech involve fixed patterns of copula, formal verbs and auxiliary verbs. Particles that consist of multiple morphemes and function like a single particle are common as well.

In our experiment, words that belong to particle, copula or auxiliary verb, were regarded as function words. Bigrams of function word were extracted from the training corpus based on the frequency $F(w_1, w_2)$ to yield function word string lists of six different sizes. Each of the selected word strings was used as a new recognition unit. Then the function word string language models FNC25, 50, 100, 151, 200, 355 was generated in the same manner as LLR LMs. Test set perplexity by the FNC LMs showed slight improvement in all conditions of the extended lexicon size (Table 2).

3.2.2 Short word string (SHT)

There are 915 lexical entries in the baseline dictionary whose length are less than three phonemes; 851 of them are non-function words. These very short words can also be regarded as potential candidates sensitive to noise in speech recognition stage. Performance improvement may be expected by adopting n-grams of the short words as longer units for recognition.

Table 1 is the number of n-gram entries composed of up to five words whose length is shorter than three phoneme extracted from the training corpus. All the candidate n-grams whose frequency is grater than 34 were first sorted by the decreasing order of frequency $F(w_1, w_2, ..., w_n), n = 2,3,4,5$ regardless of the value of n to make a single list. Then the six short word string language models of different extension vocabulary size were generated by taking the top 500, 1000, 1500, 2000, 2500, and all (3151) word strings in the list.

Test set perplexity by SHT models in any lexicon size was almost the same as FNC models, again giving negligible reduction compared to baseline and slight increase from the LLR case (Table 2).

Table 1: Short word string entries of SHT models of various sizes

size	2gram	3gram	4gram	5gram
SHT500	323	118	47	12
SHT1000	611	240	109	40
SHT1500	883	379	165	73
SHT2000	1148	514	232	106
SHT2500	1416	661	298	125
SHT3151	1719	880	387	165

4 Speech recognition experiment

4.1 Decorder

An LVCSR system was implemented with HTK [11]. A standard acoustic analysis method was employed. Each frame of input speech was represented by a 38-dimensional feature vector that consists of 12 MFCC parameters and their differential coefficients of 1st and

Table 2: Test set perplexity (trigram) of word string language models with various sizes of additional lexical entries of word string

Langage Model	Lexicon Size	Test Set PP
Baseline	18.8k	33.24
LLR500	18.8k+500	31.12
LLR1080	18.8k+1080	26.91
LLR2160	18.8k+2160	33.59
FNC25	18.8k+25	32.90
FNC50	18.8k+50	32.67
FNC100	18.8k+100	32.90
FNC151	18.8k+151	32.90
FNC200	18.8k+200	32.90
FNC355	18.8k+355	32.90
SHT500	18.8k+500	32.22
SHT1000	18.8k+1000	32.45
SHT1500	18.8k+1500	32.67
SHT2000	18.8k+2000	32.67
SHT2500	18.8k+2500	32.67
SHT3151	18.8k+3151	32.90

2nd order, together with the differential coefficients of 1st and 2nd order of speech power.

Gender-dependent triphone sets were created using ATR Phonetically Balanced sentences spoken by 201 male and 203 female speakers. The read speech materials were gathered from ATR, ASJ, and JNAS databases. The acoustic context was modeled by a shared-state, word-internal triphone model; 1294 triphone models were generated for male, and 1177 models for female; the number of Gaussian mixture components for each state is 16. Speaker and noise adaptation by MLLR was then performed on each gender-dependent model using the news speech data with various noisy environments, to generate announcer-dependent noisy speech models, which improved the baseline recognition performance by 4.5 points to give 89.9% word accuracy.

The decorder was run in the first pass performing 300-best recognition, using the bigram language models, with the fixed beam width of 250 followed by trigram rescoring of the second pass to give 50-best results.

4.2 Results

Speech recognition performance were evaluated by word correct, word accuracy [1], and deletion, substitution, insertion rate [2].

[1] $WordAccuracy[\%] = \frac{N-D-S}{N} \times 100\%$, $WordCorrect[\%] = \frac{N-D-S-I}{N} \times 100\%$, where N is the total number of word tokens, D: deletion errors, S: substitution errors, and I: insertion errors.

[2] Calculated as the percentage of each errors to the total number of word tokens, N.

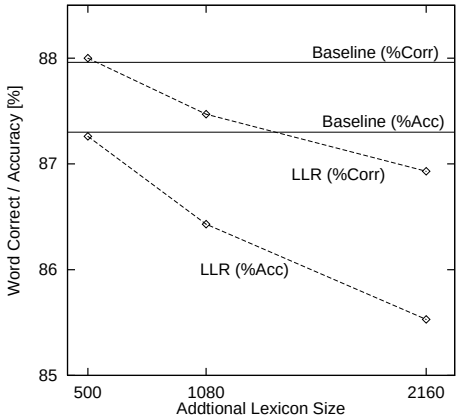

Figure 1: Noisy speech recognition performance by LLR as a function of additional lexical entry size

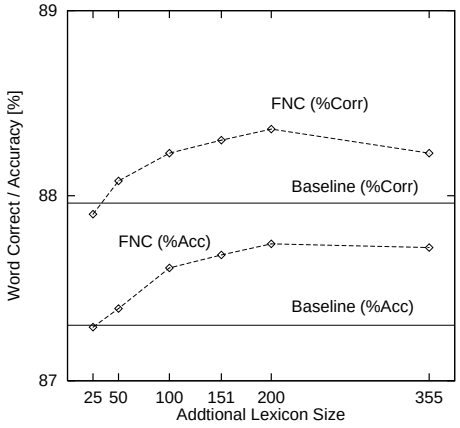

Figure 2: Noisy speech recognition performance by FNC as a function of additional lexical entry size

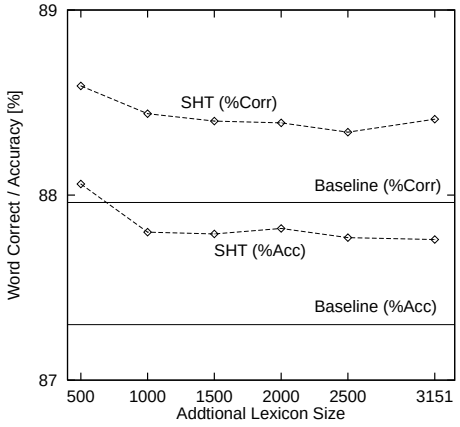

Figure 3: Noisy speech recognition performance by SHT as a function of additional lexical entry size

Table 3: Speech recognition performance of the word string language models

lm	Word Accuracy %			Word Correct			Deletion %		Substitution %		Insertion %	
	clean	noisy	total	clean	noisy	total	clean	noisy	clean	noisy	clean	noisy
Baseline	92.73	87.96	90.52	92.09	87.30	89.88	1.75	3.61	5.52	8.42	0.63	0.67
LLR500	91.52	88.00	89.89	90.74	87.26	89.13	2.12	3.49	6.36	8.50	0.78	0.74
FNC200	92.65	88.36	90.66	91.82	87.74	89.93	1.84	3.51	5.15	8.13	0.82	0.54
SHT500	92.47	88.59	90.67	91.52	88.06	89.91	1.81	2.98	5.72	8.42	0.95	0.53

Focusing attention on the noisy speech data, word string language models selected by POS and frequency criteria, i.e., SHT and FNC outperformed LLR model that is based on log likelihood selection. Both word correct and accuracy of SHT and FNC were better than those of the baseline model except for the FNC25 case, while in LLR case the performance was found to degrade as more word strings are added to lexicon as presented in Figure 1 despite the advantage in test set perplexity measure. In FNC case, word correct and accuracy improved as more strings were added to lexicon. It exhibited the best performance by FNC200 case and slight degradation by FNC355 (Figure 2). As for SHT language model, the improvement was larger than FNC. The best performance was obtained by SHT500 (Figure 3).

The best speech recognition results of each word string language models are presented in Table 3. Although the overall word correct and accuracy improvements by word string models are not significant, their advantage in deletion, substitution, and insertion errors in noisy condition is obvious. The deletion error by baseline model was reduced by 17% using SHT500 (3.61% to 2.98%). The insertion error was reduced by as much as 20% (0.67% to 0.54 or 0.53%) using FNC200 or SHT500 word string model, while substitution error improvement was 3% (8.42% to 8.13%).

An examination of recognition results showed that by the word string models based on POS and frequency criteria 97% of short function strings in the evaluation sentences were covered, while only 11% of them by the word string models based on log likelihood ratio criteria.

5 Conclusion

In this paper, we presented the results that the word string language model whose extended vocabulary was selected based on part of speech information and occurrence frequency criteria improves recognition performance of noisy broadcast news speech. The word string language models based on POS and n-gram frequency reduces deletion errors by 17%, insertion errors by 20%, and substitution errors by 3% in Japanese TV broadcast news speech recognition. The result provides an approach from language model side to remedy against recognition errors of short words.

Acknowledgments

This work was partially funded by, and the broadcast news corpus was provided by NHK (Japan Broadcasting Corporation) Science and Technical Research Laboratories.

References

[1] D. Pallett, J. Fiscus, J. Garofolo, A. Martin, and M. Przybocki, "1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures," Proc. DARPA Broadcast News Workshop, pp. 5 - 12, Feb. 1999.

[2] A. Ando and E. Miyasaka, "Construction of Japanese News Speech Databases," Proc. Acoustical Society of Japan Spring Meeting, 2-Q-9, Mar. 1997.

[3] K. Ohtsuki, S. Furui, N. Sakurai, A. Iwasaki, and Z. Zhang, "Improvements in Japanese Broadcast News Transcription," Proc. DARPA Broadcast News Workshop, pp. 231 - 236, Feb. 1999.

[4] T. Imai, A. Kobayashi, S. Sato, H. Tanaka, and A. Ando, "Progressive 2-Pass Decorder For Real-Time Broadcast News Captioning," Proc. ICASSP2000, Vol.3, pp. 1559 - 1562, Jun. 2000.

[5] B. Suhm and A. Waibel, "Towards Better Language Models for Spontaneous Speech," Proc. ICSLP 94, S16-4, pp. 33 - 38, Sept. 1994.

[6] T. Kobayashi, Y. Wada, and N. Kobayashi, "Source-Extended Language Model for Large Vocabulary Continuous Speech Recognition," Proc. ICSLP98, Session Th4P13, Dec. 1998.

[7] D. Klakow, X.Aubert, P. Beyerlein, R. Haeb-Umbach, M. Ullrich, A. Wendemuth, and P. Wilcox, "Language-Model Investigations Related to Broadcast News," Proc. DARPA Broadcast News Transcription and Understanding Workshop, Feb. 1998.

[8] ChaSen (Japanese Morphological Analysis System) Version 2.0, http://cl.aist-nara.ac.jp/lab/nlt/chasen/

[9] P. Clarkson, R. Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit," Proc. Eurospeech97, pp. 2707 - 2710, Sept. 1997.

[10] M. Utiyama, "Phrase Addition for Large Corpora," Proc. Acoustical Society of Japan Fall Meeting, 2-1-14, Sep. 1998.

[11] HTK: Hidden Markov Model Toolkit, Version 2.01, http://www.entropic.com/