BENCHMARK TEST FOR SPEECH RECOGNITION USING THE CORPUS OF SPONTANEOUS JAPANESE

Tatsuya Kawahara *Hiroaki Nanjo* *Takahiro Shinozaki* *Sadaoki Furui*

School of Informatics,
Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

Department of Computer Science,
Tokyo Institute of Technology
Meguro-ku, Tokyo 152-8552, Japan

ABSTRACT

We present benchmark results of automatic speech recognition using the Corpus of Spontaneous Japanese (CSJ), which has been developed in the five-year national project and will be the largest spontaneous speech databases. New test-sets are designed for both academic presentation speech and extemporaneous public speech, which are the two major categories in the corpus. The test-sets are selected to cover the variation of acoustic and linguistic factors in spontaneous speech: word perplexity, degree of disfluency, and the speaking rate. Baseline acoustic and language models are set up using an almost complete set (500 hours and 6.67M words) of the CSJ. Statistical modeling of pronunciation variation is also incorporated into the language model based on the alignment of large-scale transcriptions. The benchmark results verified the effects of the factors considered in the test-set design.

1. INTRODUCTION

Automatic speech recognition (ASR) of read speech has successfully achieved accuracy exceeding 90% and realized a dictation system. The system, however, assumes that users clearly utter grammatically correct sentences with orthodox pronunciation for human-to-machine interfaces. On the other hand, recognition of human-to-human spontaneous speech, which would make possible automatic transcription or translation of lectures and meetings, is very poor and needs more extensive studies.

From this perspective, the five-year project "Spontaneous Speech Corpus and Processing Technology" has been conducted since 1999[1]. The foremost product of the project is a large-scale spontaneous speech corpus[2]. The *Corpus of Spontaneous Speech (CSJ)* consists of roughly seven million words. Monologues such as lectures, oral presentations and extemporaneous speeches are mainly recorded. Compared with other corpora of spontaneous speech such as the Switchboard corpus, the scale of the CSJ is prominent. Another characteristic of the CSJ is that speech is input by a head-set microphone and digitally sampled at 16 kHz. Thus, we do not have to be much concerned about compensation for noise and channel distortion which are other causes of degradation in telephone conversation speech, and we can focus on the issues caused by spontaneity of speaking.

As many previous studies point out, various factors in spontaneous speech affect ASR performance. They include acoustic variation caused by fast speaking and imperfect articulation, and linguistic variation such as colloquial expressions and disfluencies. Thus, the problems have been addressed from the viewpoint of acoustic modeling, pronunciation modeling and language modeling.

With the huge scale of the CSJ, it is possible to conduct more comprehensive studies by systematically designing evaluation tests and investigating the effect of individual factors and methods. In this paper, we present ASR benchmark test-sets we designed using the CSJ and dry-run recognition results. The platform design involves construction of baseline acoustic and language models.

2. FACTORS CONSIDERED IN TEST-SET DESIGN

A large portion of the CSJ consists of two styles of monologues. One is academic presentation speech at technical conferences and meetings, and the other is extemporaneous public speech on given topics such as hobbies and travels. Since the speaking style and vocabulary are apparently different for these two categories, we set up respective test-sets. In addition, considering the fact that most of the academic presentations are given by male speakers, we set up two sets for the academic category: a male-only set and a gender-balanced set. Actually, so far we have made only the male-dependent evaluation by using another test-set[3][4][5] that is different from those presented in this paper. Thus, we have three test-sets, each of which consists of ten talks.

In the design of the test-sets, talk samples are chosen so that the sets well represent the whole corpus with respect to various factors of spontaneous speech. Shinozaki and Furui[5] investigated the correlations of various factors with speech recognition accuracy. They concluded that the speaking rate (SR), out-of-vocabulary (OOV) rate and self-repair rate (RR) are directly correlated with accuracy. Other factors are mainly dependent on either of these three. For example, word perplexity (PP) is correlated with the OOV rate and, by disregarding the correlation, PP is not so correlated with the accuracy. Similarly, the filler rate (FR) and acoustic likelihood (AL) are dependent on the speaking rate (SR).

Therefore, the three factors SR, OOV, and RR should be taken into account in the test-set selection. Among those, the out-of-vocabulary (OOV) rate is highly dependent on the vocabulary in nature and is easily variable when the lexicon is modified. So, we adopt word perplexity (PP) instead of OOV because perplexity, especially its difference among speech samples, is generally more stable even when the language model is revised.

Statistics of these features are listed in Table 1 and Table 2 for academic presentations and extemporaneous speech of the current CSJ[1], respectively. Here, the speaking rate (SR) is defined as the average number of morae per second in an utterance. The filler rate (FR) and self-repair rate (RR) are average occurrences of fillers and self-repairs [2] divided by the number of words in a talk. It is observed that speaking in academic presentations is faster and more disfluent, specifically it has more fillers and self-repairs. For reference, in the JNAS read speech corpus, the speaking rate is slower (7.36 mora/sec.) and it obviously contains no fillers and self-repairs.

Among the three major features (PP, SR, RR) defined above, PP represents linguistic difficulty, SR shows acoustic difficulty, and RR is a measure of disfluency that affects both acoustic and linguistic aspects.

3. TEST-SET SELECTION PROCEDURE

Using the statistics, we select test-set speeches so that they represent the whole corpus in terms of the three features discussed above. First, we select two talks of low word perplexity (PP) and two of high PP. Here, judgement of low and high is based on standard deviation (SD). For example, if a value is smaller than the mean minus SD, it is regarded as low. Then, six talks are chosen from samples of normal PP. Among them, we select one with a low self-repair rate (RR), one with high RR. The remaining four are chosen from samples of normal RR, such that one has a slow speaking rate (SR), one fast SR and two normal SR.

Gender balance is also taken into account except in the male-only test-set. Very long talks (longer than 30 minutes) are avoided so that all samples have similar influence on the final evaluation measure. Thus, we set up three test-sets. [3] The ID list of the test-sets is given in Table 4 and Table 5.

Table 1. Statistics of academic presentation speech (865 talks)

	mean	SD
PP	81.1	25.3
OOV (%)	1.45	0.66
SR (mora/sec)	9.05	1.09
FR (%)	6.80	3.44
RR (%)	1.40	0.79

Table 2. Statistics of extemporaneous public speech (1504 talks)

	mean	SD
PP	82.4	25.2
OOV (%)	1.71	0.82
SR (mora/sec)	7.97	0.79
FR (%)	5.45	3.25
RR (%)	1.25	0.85

4. BASELINE SYSTEM

4.1. Language Model

A baseline language model is constructed using the transcriptions of 2592 talks excluding the test-set. The total text size is about 6.67 million words[4] including fillers and word fragments. Word segmentation was automatically done using a morphological analyzer that was trained with the maximum entropy criterion by Uchimoto et al[6].

In spontaneously spoken Japanese, pronunciation variation is so large that a number of baseform entries are needed for a lexical item. We found that statistical modeling of pronunciation variations integrated with the language modeling was effective in suppressing false matching of less frequent entries[7]. Here, we adopt a simple trigram model of word-pronunciation entries.

Transcription of the CSJ was made manually both in an orthographic notation and a phonetic (*kana*) one for each utterance unit. Thus, automatic alignment of the two by the word unit is needed to obtain the word-pronunciation entries. This was incorporated as a post-processor of the morphological analyzer[6]. Some heuristic thresholding is applied to eliminate erroneous patterns. As a result, we get 30820 word-pronunciation entries, for which a trigram model is trained.

[1] As of November 2002

[2] Strictly speaking, they are tagged as word fragments, which are usually signs of self-repairs.

[3] For practical purposes, selection is made from talks included in the sample version of the CSJ, which is already publicly available.

[4] A system of short word unit defined in the project

Table 3. List of acoustic models

model		training data	
		#talks	size (hour)
academic	male	787	186
presentation	female	166	42
speech	GID	953	228
extemporaneous	male	721	124
public	female	822	134
speech	GID	1543	258
mixed	male	1508	310
	female	988	176
	GID	2496	486

GID: gender-independent model

4.2. Decoder

Julius rev.3.3p3[5] is used as a recognition engine. Sequential decoding[3] is applied so that very long speech segments can be handled without prior segmentation.

4.3. Acoustic Model

We have set up a variety of baseline acoustic models. Since the speaking style is apparently different for academic presentation speech and extemporaneous public speech, respective models are trained in addition to the mixed model that uses the whole available data except the test-set. For each category, both gender-dependent and gender-independent models are prepared.

The list of the acoustic models with their training data sizes is given in Table 3. The training data of male speakers is much larger for the academic presentations. All the models are triphone HMMs that have 3000 shared states with 16 Gaussian mixture components.

5. RECOGNITION RESULTS

Word accuracies for academic presentation speech (test-sets 1 and 2) are given in Table 4 and those for extemporaneous public speech (test-set 3) are in Table 5.

For academic presentations, the gender-independent model trained only with speech samples of the same style achieves the best performance on average. The gender-dependent models give slightly lower accuracy. Large degradation is observed for some specific speakers such as A01M0110 and A03F0072.

As for extemporaneous speech, the model trained with all available data including academic presentations obtains the best accuracy. This fact suggests that the speaking style of extemporaneous speech is more general than that of academic presentations. In this case, the gender-dependent model is better than the gender-independent model except for one speaker (S00F0148).

Next, we investigate the correlations of the word error rate (WER) with the three major factors that were taken into account in selecting the test-sets. They are plotted in Fig. 1, Fig 2 and Fig 3.

The effects of the word perplexity and speaking rate are confirmed. By limiting to academic presentations, their correlation coefficients are much larger (0.53 and 0.36). However, the correlation with the self-repair rate is not observed unlike the previous study[5] in either case. This suggests that disfluency can be a very complex phenomenon.

6. CONCLUSIONS

New test-sets for speech recognition using the CSJ were presented. They were carefully designed by considering the major factors in spontaneous speech: word perplexity, degree of disfluency and the speaking rate.

Baseline models are also set up using all currently available data of the CSJ, which has been almost completed in the five-year project. Benchmark results with the baseline system were presented for the test-sets. Different tendencies were observed in academic presentations and extemporaneous speech. Correlations of the accuracy with the word perplexity and speaking rate are confirmed. More careful investigations should be done to make clear the problems and future directions of spontaneous speech recognition.

Acknowledgment: The work was conducted in the Science and Technology Agency Priority Program on “Spontaneous Speech: Corpus and Processing Technology”. The authors are grateful to Dr. Yasuhiro Minami, Dr. Kiyotaka Uchimoto, and Dr. Atsushi Yamada and other members of this productive project for their collaboration.

REFERENCES

[1] S.Furui. Recent advances in spontaneous speech recognition and understanding. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (this volume), 2003.

[2] K.Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (this volume), 2003.

[3] T.Kawahara, H.Nanjo, and S.Furui. Automatic transcription of spontaneous lecture speech. In *Proc. IEEE workshop on Automatic Speech Recognition and Understanding,* 2001.

[4] H.Nanjo and T.Kawahara. Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. In *Proc. IEEE-ICASSP*, pages 725–728, 2002.

[5] T.Shinozaki and S.Furui. Analysis on individual differences in automatic transcription of spontaneous presentations. In *Proc. IEEE-ICASSP*, volume 1, pages 729–732, 2002.

[5] http://julius.sourceforge.jp/

[6] K.Uchimoto, C.Nobata, A.Yamada, S.Sekine, and H.Isahara. Morphological analysis of Corpus of Spontaneous Japanese. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (this volume), 2003.

[7] H.Nanjo and T.Kawahara. Unsupervised language model adaptation for lecture speech recognition. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (this volume), 2003.

Table 4. Word accuracy for academic presentation speech

	acoustic model			
	academic		academic+extempo.	
(gender) ID	GD	GID	GD	GID
test-set 1				
(m) A01M0097	86.88	86.72	83.91	83.10
(m) A04M0051	81.59	81.98	80.77	80.73
(m) A04M0121	70.54	71.51	70.45	69.74
(m) A03M0156	58.07	57.35	56.40	55.29
(m) A03M0112	83.02	82.31	82.75	80.93
(m) A01M0110	71.77	82.40	70.24	82.17
(m) A05M0011	67.10	67.03	68.77	67.05
(m) A03M0106	59.16	60.49	59.21	61.55
(m) A01M0137	73.97	73.73	73.97	72.71
(m) A04M0123	70.89	70.30	68.18	68.78
test-set 1 total	71.90	72.45	71.12	71.24
test-set 2				
(f) A01F0063	69.42	55.95	67.29	54.19
(m) A01M0056	85.26	84.70	84.96	83.98
(f) A06F0135	81.97	78.27	80.82	79.50
(m) A02M0012	74.54	74.13	74.84	73.95
(m) A06M0064	67.08	67.08	70.04	67.36
(m) A01M0141	80.03	79.47	78.60	76.86
(f) A01F0034	80.18	77.78	80.08	78.87
(m) A03M0016	63.93	60.55	63.71	60.97
(f) A03F0072	50.02	71.52	51.51	70.26
(f) A01F0001	77.20	77.08	75.38	76.09
test-set 1+2 total	72.03	72.58	71.58	71.78

Table 5. Word accuracy for extemporaneous public speech

	acoustic model			
	extempo.		academic+extempo.	
(gender) ID	GD	GID	GD	GID
test-set 3				
(f) S00F0066	75.15	72.00	76.53	71.63
(m) S00M0213	83.33	83.39	83.67	82.93
(m) S00M0070	83.29	80.74	84.47	81.36
(m) S00M0008	64.14	61.55	64.34	62.99
(f) S01F0105	76.98	76.83	79.25	76.75
(f) S00F0148	55.57	63.26	56.01	63.20
(f) S00F0019	81.61	79.65	81.68	78.22
(m) S00M0112	67.26	67.23	71.56	70.56
(f) S00F0152	67.14	63.74	68.66	62.57
(m) S00M0079	69.81	66.70	70.03	66.58
test-set 3 total	71.92	71.07	73.27	71.46

GD: Gender-Dependent, GID: Gender-Independent

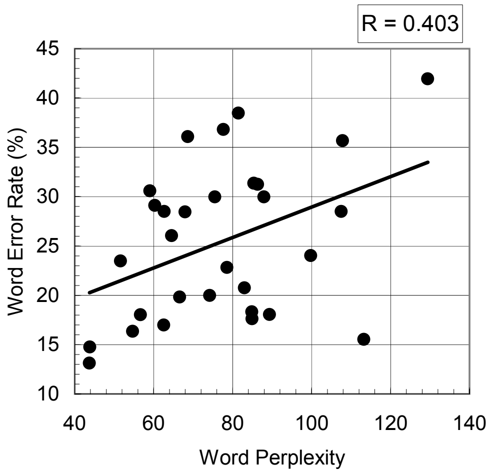

Fig. 1. Relation of word perplexity (PP) and word error rate

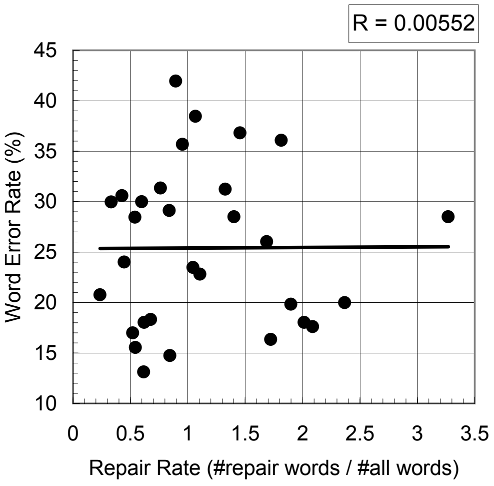

Fig. 2. Relation of self-repair rate (RR) and word error rate

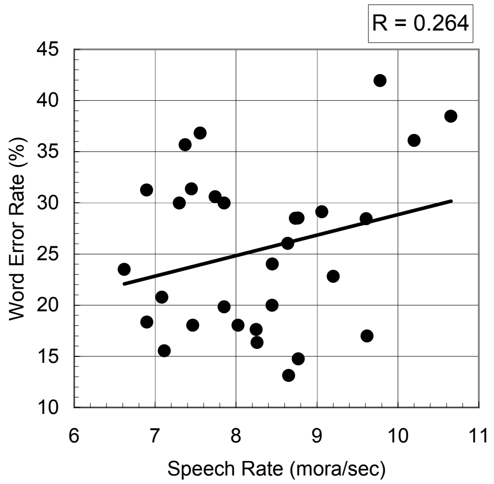

Fig. 3. Relation of speaking rate (SR) and word error rate