interspeech 2008 incorporating SST 08

Speech Recognition Performance of CJLC: Corpus of Japanese Lecture Contents

Satoru Kogure[1], *Hiromitsu Nishizaki*[2], *Masatoshi Tsuchiya*[3], *Kazumasa Yamamoto*[4]
Shingo Togashi[4], *Seiichi Nakagawa*[4]

[1] Faculty of Informatics, Shizuoka University

[2] Department of Interdisciplinary Research Graduate School of Medicine and Engineering, University of Yamanashi

[3] Information and Media Center, Toyohashi University of Technology

[4] Department of Information and Computer Sciences, Toyohashi University of Technology

kogure@inf.shizuoka.ac.jp, hnishi@yamanashi.ac.jp
{tsuchiya,kyama,togashi,nakagawa}@slp.ics.tut.ac.jp

10.21437/Interspeech.2008-444

Abstract

This paper discusses the speech recognition of Japanese classroom lecture speech. In particular, we mention the influences of microphone differences and the language model differences on the speech recognition performance of classroom lectures. First, we collected actual classroom lecture contents from several universities in Japan. In this paper, we recorded the lecture speech using lapel microphones because lapel microphones are more commonly used to record lectures. LVCSR is one of the essential technologies for adding tag information to such lecture speech. Next, therefore, we researched the influence of the differences between microphones used for recording lecture on speech recognition performance. Finally, seven types of language models that were trained using three types of corpora were compared on the basis of their ability to lecture speech.

Index Terms: classroom lecture contents, lecture speech recognition, spontaneous speech, lapel microphone, language model

1. Introduction

Recently, e-learning systems for the preparation and review of lectures, that is, browsing, have been developed and most of these systems are available in the field of education today. There exists considerable commercial or free software of e-learning systems or learning management systems (LMSs), such as eCampus [1], IT's class [2], Blackboard [3], and so on.

In addition, there is commercial e-learning software that can create lecture contents and present a lecture simultaneously. For instance, "EZ presentator" developed by Hitachi Advanced Digital Inc. can create an e-learning content that include a video of a lecture from a Microsoft PowerPoint ® file. The contents can be viewed through an Web browser. When a user chooses a slide, the part of the video that explains the slide can be played since the turn timing of each slide is recorded and the video is synchronized with each slide turning.

However, users cannot directly access the points that they want to watch in most e-learning systems such as "EZ presentator." In other words, although they can search slides by using the titles, they cannot search slides by the contents. Therefore, it is necessary to develop e-learning systems that intelligently process the contents of lectures. The most important and basic technology for doing so is robust speech recognition of spontaneous speech. Furthermore, spoken document retrieval [1], automatic speech summarization [2, 3] techniques, etc. , may improve the usability of the systems and should be the essential key elements. These technologies assume using an output from an automatic speech recognizer.

However, there are various problems related to recognizing classroom lecture speech. Recently, some projects related to classroom lecture speech have been carried out [4, 5, 6, 7].

For instance, **CSJ** (Corpus of Spontaneous Japanese) has been published in Japanese [8] and is used by many researchers. This is a large speech corpus of about 3,300 Japanese spontaneous speech lectures, including a thousand lectures presented at Japanese academic meetings and about 1,600 simulated lectures. As each speech included in CSJ was recorded with a headset microphone, the recorded speech is clear as compared to the speech recorded with a lapel microphone. One of our final goals is to achieve an improvement in the speech recognition performance of classroom lecture speech recorded with a lapel microphone.

Therefore, we have collected considerable Japanese classroom lecture speech at a couple of universities for developing technologies of robust speech recognition and for advanced processing the lecture contents. Furthermore, we will release this classroom lecture database for research usage. We call it the "Corpus of Japanese classroom Lecture Contents" (abbreviated as **"CJLC"**). We hope that this corpus helps in making a breakthrough in the technologies of spoken language processing used by many researchers.

This paper investigates the basic recognition performances of classroom lecture speech recorded with lapel microphones included in CJLC. We think that it is difficult for teachers to use a handheld microphone because its use may hamper the progression of the lecture, especially in the case of using a blackboard. Therefore, when classroom lecture speech is recorded, a lapel microphone may be used instead of a handheld one. Further, the classroom lecture speech recorded with a lapel microphone includes various noises such as the babble of students and reverberant speech. We must face these challenges such that the spontaneous speech recorded with a lapel microphone is robustly and automatically transcribed. Therefore, in this paper, we have compared the recognition performance of speech recorded with various microphones. We used four microphones (a high-performance lapel, a normal-performance lapel a normal-performance handheld, and a normal-performance headset microphones) to record lecture speech.

For speech recognition, the manner in which a language model is developed is also important. In this paper, seven kinds of language models that were trained using the transcriptions of CSJ, newspaper articles, and Web corpus were developed and used to recognize the lecture speech of CJLC. Experimental results showed that the CSJ language models is quite well compared with other language models; however, they were not sufficient for text processing. There is still room for improve-

[1] http://excampus.nime.ac.jp/index.html
[2] http://www.gp.hitachi.co.jp/eigyo/ product/itsclass/
[3] http://www.blackboard.com/us/index.Bb

Table 1: Lecture speech used for recognition experiments.

Lec.ID	Spk.ID	#Snt	Time [s]	Filler Rate [%]	Class	Keyword for Lecture Contents
L1	S1	259	1213	6.80 (263/3869)	LS1: Computer Applications II	Spoken Language Processing
L2		236	1274	4.54 (179/3946)		DP Matching
L3		212	160	4.70 (80/1702)		Language Model
L4	S2	668	937	14.75 (496/3364)	LS2: Computer Applications I	Natural Language Processing
L5		311	254	10.09 (187/1853)		
L6	S3	1480	3623	6.64 (1006/15157)	LS3: Software Engineering	Design of Program, coding
L7	S4	743	1903	8.20 (484/5901)	LS4: Experiments on Physics	Diode, P-type and N-type Semiconductor
L8	S5	1163	4193	5.69 (672.11803)	LS5: Algorithm and Data Structure I and Practice	Binary Tree II
L9		903	3115	6.57 (550/8367)		Bubble Sort
L10		820	3285	8.24 (745/9037)		Selection and Insertion Sort
L11		564	2261	7.72 (504/6529)		Quick Sort
CSJ		1771	4568	10.91 (2050/18793)	CSJ	The Acoustical Society of Japan, etc.

ment. Therefore, we also focus on a language model adaptation for classroom lecture speech. The recognition performance of lecture speech improves when language models trained using the transcriptions of a lecture related to the target lecture are used. Hence, we compiled a training corpus by searching Wikipedia for Japanese text using the keyword sets extracted from the target lecture's syllabus information and trained the language models for lecture speech recognition using a combination of the training corpus and CSJ. As a result, the word accuracy improved by 2.4%.

2. Overview of CJLC

CJLC is formally defined as a set of classroom lecture data, and each data element includes "a lecture speech," "its synchronized transcription," "a presentation slide data" (optional), "a timetable of slide show" (optional), and "a list of important utterances" (optional). The numbers of speakers, courses, and lectures are 15, 26, and 86, respectively. Further, the total duration is 3,780 min.

Further, we compared the classroom lectures of CJLC and the lecture of CSJ. The result of analysis showed that the average number of filled pauses included in the classroom lectures was almost the same as that in CSJ. On the other hand, the average number of word fragments caused by disfluency acts in the classroom lectures was less than that in CSJ speech. The details are given in [10].

3. LVCSR Performance on CJLC speech

We have collected classroom lectures from real environments under various conditions such as different microphones for speech recognition experiments.

We investigated how the differences between microphones affect the performance of speech recognition. Classroom lecture speech was recorded with four types of microphones. We also investigated the differences in recognition performances between language models. We used seven different language models; three types of language models trained using CSJ, two language models created using news articles and two models using Web collections for the experiments.

Table 1 shows the details of the lecture speech which was selected from CJLC. Various lectures from five courses were prepared.

We used Julius rev.3.5.3 for speech recognition. This is an open source decoder for LVCSR and runs in two decoding passes; the first pass uses a word bigram and the second pass uses a word trigram. Based on the word trigram and context-dependent HMM, Julius can perform real-time decoding on a 20k vocabulary dictation task in most of the current PCs.

Julius used triphone-based HMMs trained using CSJ recorded with a high-quality headset microphone (CROWN CM-311A), which were sampled at 16 KHz and 16 bits. Feature vectors comprised of 38 dimensions: 12 dimensional Mel-frequency cepstrum coefficients (MFCCs), the cepstrum difference coefficients (delta MFCCs), their acceleration (delta delta MFCCs), delta power, and delta delta power; these vectors were calculated every 10 ms. The distributions of the acoustic features were modeled using 32 mixtures of diagonal covariance Gussians for the HMMs.

Table 2: Training conditions of language models.

LM ID	Vocab.	Training Data	
		# of Lectures	Size [Byte]
$CSJ_{970-20k}$	20k	970 *1	23 MB
$CSJ_{3300-20k}$	20k	3285 *2	123 MB
$CSJ_{3300-40k}$	40k		
$NEWS_{20k}$	20k	1,499,936*3	1400 MB
$NEWS_{42k}$	42k		
WEB_{20k}	20k	—*4	100 GB
WEB_{60k}	60k		

*1: 970 real lectures at Japanese academic meetings
*2: 970 real lectures at Japanese academic meetings + 1715 simulated lectures + 523 readings + 58 dialogs and 19 etc.
*3: all articles from 1991 to 2004 on Mainichi newspaper, Japan.
*4: The LM was prepared by Fujii *et al.* [9] and trained using the 100G text of Web collection.

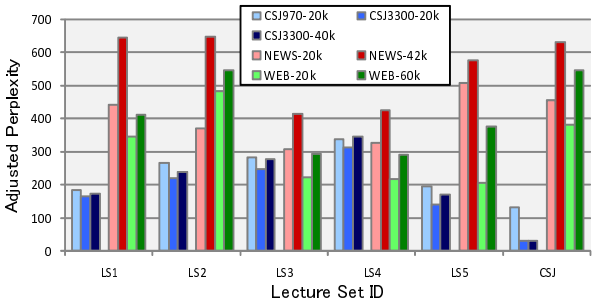


Figure 1: Test set adjusted perplexity based on different language models.

Table 2 shows the training conditions of each language model. Three types of models based on CSJ, which differed in the number of lectures of training and vocabulary size, were used in the decoder. Two types of models based on Mainichi newspaper articles, that differed only in vocabulary size, were used. Furthermore, two types of models that had a large vocabulary of 20k and 60k, respectively, were trained using Web articles.

3.1. Influence of microphone performance

In order to test the influence of microphone performance or wired/wireless condition, we recorded the two classroom lec-

Table 3: Word recognition rates of lectures recorded with four microphones[%].

(AM: Triphone / LM: $CSJ_{3300-40k}$)

mic.	(a) normal lapel	(b) high lapel	(c) normal handheld	(d) normal headset
Acc.[%]	55.4	56.4	60.0	62.7
Cor.[%]	61.3	62.7	67.3	70.5

tures, L3 and L5 in Table 1, with four types of microphones. The four microphones are (a) SONY ECM-C10 (normal/lapel), (b) SONY ECM-88B (high/lapel), (c) SONY ECM-355 (normal/handheld), and (d) ISOMAX Headset Microphone (normal/headset).

In this experiment, $CSJ_{3300-40k}$, as represented in Table 2, was used for speech recognition. Table 3 shows the results of the experiment on the average of L3 and L5. From these, we could state the order of recognition performance as follows: headset microphone $>$ handheld microphone $>$ lapel microphone (high performance) $>$ lapel microphone (normal performance). In order to get a higher recognition rate for lecture speech, a higher-quality microphone should be used to record speech.

3.2. Language models

As described in Table 2, we prepared the seven language models and evaluated them on test set perplexity and speech recognition performance.

We calculated the *test set adjusted perplexity* (APP), which was given by adjusting the PP for taking account of OOV words [11], for five lecture sets (from LS1 to LS5) and the CSJ test set given in Table 1. The values of APP and rate of OOVs are shown in Fig. 1 and Fig. 2. For almost all of the lectures, the values of APP of the CSJ_{3000} model were lower than the ones of the CSJ_{970} model. This was because the utterances in usual lectures that often include short dialogue, discourse, or monology were similar to the contents of the 3,300 lectures in CSJ, whereas all the training data of the 970 lectures was based only on the lectures presented at academic meetings in Japan. Moreover, the values of APP were high in the order CSJ $>$ Web $>$ NEWS because the news articles consisted of formal sentences and the sentence style in th Web collection was casual (similar to that of CSJ) but had formal sentence structures (similar to that of NEWS). Figure 3 shows the filler rate in OOVs. Four language models, WEB-20k, WEB-60k, NEWS-20k, and NEWS-42k contain at most several dozen filler words (that is, 9, 17, 2, and 4 types of filler words, respectively). On the other hand, three CSJ language models, 970-20k, 3300-20k, and 3300-40k contain about 1,000 types of filler words. Figure 4 shows the occurrence rate of alphabet, number and loan words. The rates of number (OOV) and alphabet (OOV) for LS4 and LS5 are higher than the ones of LS1, LS2, LS3, and CSJ because the contents of LS4 and LS5 are quite different from the contents of CSJ.

Next, we evaluated these language models on the basis of the speech recognition rate of five lecture sets and the CSJ test set. Figure 5 shows the word accuracy of the five lecture sets and the CSJ lecture set. The results showed that the lecture recognition using the $CSJ_{3300-20k}$ language model had a better performance than that using the CSJ 970 language model. In particular, the difference in the recognition performance was salient in LS3, LS4, and LS5. The word accuracy using WEB-20k was about the same as that of $CSJ_{3300-20k}$. Moreover, the recognition performances using NEWS language models were determinately bad when compared to the word accuracy of CSJ or Web collection language models. These results were attributed to the fact that the filler rate in the OOVs of NEWS was considerably higher than that in the OOVs of the CSJ or Web collection language models as shown in Fig. 3. Moreover,

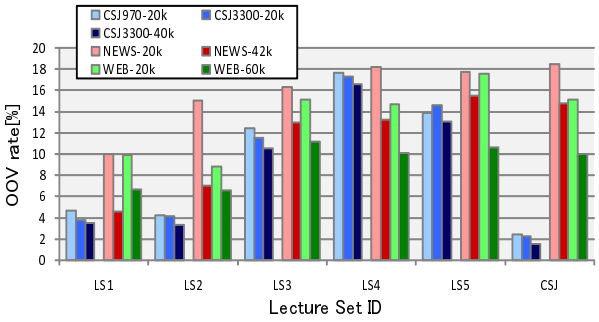


Figure 2: Rate of OOVs based on different language models.

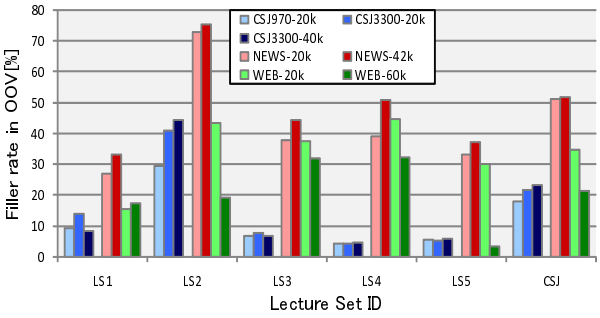
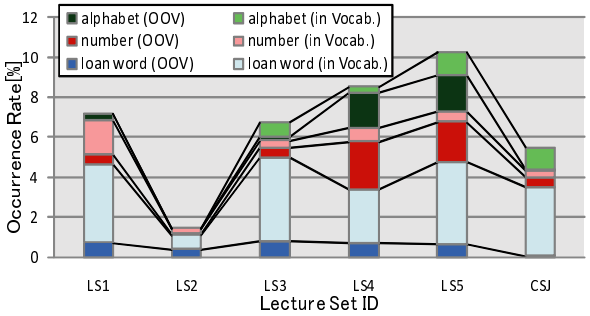


Figure 3: Filler rate in OOVs of different language models.

the filler rate of WEB-60k was lower than that of WEB-20k because the filler words that had a high frequency in lecture speech were contained in the vocabulary of the WEB-60k model but were not contained in the vocabulary of WEB-20k model.

From these results, we concluded that it was better to use the language models trained using a spontaneous or casual expression corpus than to use the language models trained using a formal/written style expression corpus for lecture speech recognition.

In order to adapt the lecture information to the language models, we used the lecture's syllabus information. We prepared the syllabus information of the target lecture in brief. When adapting the task information to the language model, we should collect the corpus related to the task. The popular manner is the method of retrieving Web using the keyword set related to the task and adapting the language model by using the



LM: $CSJ_{3300-40k}$

Figure 4: Occurrence rate of alphabet, number, and loan words

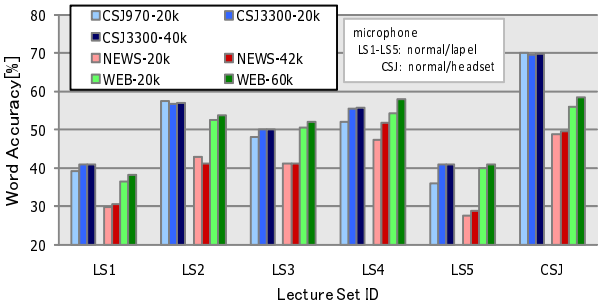
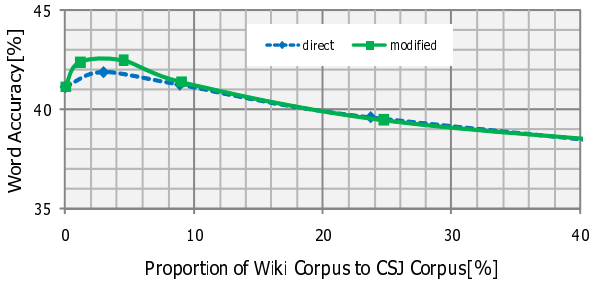


Figure 5: Word accuracy of lecture speech recognition.



(microphone: normal/lapel).

Figure 6: Word accuracy of different ratios of wiki to CSJ.

retrieved corpus. In this method, we should delete unnecessary pages from the obtained corpus because it often contains the pages that are poorly correlated to the task. However, the selection of only the necessary pages is high cost. Therefore we focused on using the Web dictionary. In the use of the Web dictionary, it is clear to be able to considerably reduce the number of pages disrelated to the task in the training corpus obtained by retrieving. Furthermore, we select Wikipedia as Web dictionary site because Wikipedia can download all information in the dictionary on ahead.

After we compiled a keyword set from the contents of syllabus, we repeated "retrieving information from Wikipedia using the prepared/new keyword sets" and "extracting the new keyword set from the retrieved data sets" twice. At the end, we compiled a wiki corpus from the results of searching the Wikipedia database using the last keyword set and trained the wiki language models using collected wiki corpus. We found and confirmed that the performance of these language models was considerable worse than that of CSJ. Therefore, we prepared a combination corpus of the wiki corpus and the CSJ and trained the language models using the combination corpus. We varied the ratio of the wiki corpus to the CSJ from 0% to 40% (0% means using only CSJ).

Figure 6 shows the word accuracy using the adapted language models in lecture set LS5. The term "Direct" implied that the second and the third keyword sets were automatically selected. The term "Modified" implied that the second and the third keyword sets were manually selected. By using the adapted language models, the word accuracy was increased by about 1.5% (error reduction rate of 2.6%), for the lecture average (up to about 2.4% in lecture L10). Moreover, by using adapted language models, the noun word correct rate was increased by about 4.7% (error reduction rate of 8.9%) in the four lecture average (up to about 6.4% in lecture L10).

From these results, we showed that an adaptation of the language model that uses Web correction was more effective in recognizing the lecture speech.

4. Conclusions

In order to study the current state of classroom lecture speech recognition, which is one of the fundamental technologies needed to process lecture contents, we recorded the lectures in real-time condition. Moreover, according to the speech collected by using various microphones, we clarified the current state of lecture speech recognition. We showed that the microphone quality had a critical influence on the accuracy rates. In the case when the content of a classroom lecture was similar to an academic lecture, language model $CSJ_{970-20k}$ had a higher performance than $CSJ_{3300-20k}$. In contrast, when the content of the classroom lecture was different from that of an academic lecture, the $CSJ_{3300-20k}$ language model had a higher performance than CSJ_{970}. In addition, by adapting the syllabus information to the language model, we clarified that the word accuracy and the noun word correct rate could be improved.

The monitor version of CJLC is already available. Please see http://www.slp.ics.tut.ac.jp/CJLC/. We will release the formal version of the CJLC database to the public for research use only in the near future.

5. Acknowledgments

This research was supported by Strategic Information and Communications R&D Promotion Programme of Ministry of Internal Affairs and Communications. Furthermore, we also thank teachers for their cooperation to record the classroom lecture speech.

6. References

[1] Hiromitsu Nishizaki and Seiichi Nakagawa, "Japanese spoken document retrieval considering OOV keywords using LVCSR system with OOV detection processing," in *Proc. of Human Language Technology Conference(HLT)2002*, 2002, pp. 144–151.

[2] Chiori Hori, Takaaki Hori, and Sadaoki Furui, "Evaluation method for automatic speech summarization," in *Proc. of the 8 th European Conference on Speech Communication and Technology(EUROSPEECH'03)*, 2003, pp. 2825–2828.

[3] S. Togashi, M. Yamaguchi, and S. Nakagawa, "Summarization of spoken lectures based on linguistic surface and prosodic information," in *Proc. of the IEEE/ACM Workshop on Spoken Language Technology(SLT)*, 2006, pp. 34–37.

[4] A. Park, Timoty J. Hazen, and James Glass, "Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling," in *Proc. of 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2005)*, 2005, pp. 497–500.

[5] James Glass, Timothy J. Hazen, Lee Hetherington, and Chao Wang, "Analysis and processing of lecture audio data: Preliminary investigations," in *Proc. of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*, 2004, pp. 9–12.

[6] L.Lamel, E.Bilinski G. Adda, and J.L. Gauvain, "Transcribing lectures and seminars," in *Proc. of the 9^{th} European Conference on Speech Communication and Technology(EUROSPEECH2005)*, 2005, pp. 1657–1660.

[7] Isabel Trancoso, Ricardo Nunes, Luís Neves, Céu Vianan, Helena Moniz, Diamonino Caseiro, and Ana Isabel Mata, "Recognition of Classroom Lectures in European Portuguese," in *Proc. of the 9^{th} International Conference on Spoken Language Processing (Interspeech2006-ICSLP)*, 2006, pp. 281–284.

[8] Kikuo Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, 2003, pp. 7–12.

[9] A. Fujii and K. Itoh, "Building a test collection for speech-driven web retrieval," in *Proc. of EUROSPEECH2003*, 2003, pp. 1153–1156.

[10] Masatoshi Tsuchiya, Satoru Kogure, Hiromitsu Nishizaki, Kengo Ohta and Seiichi Nakagawa, "Developing corpus of Japanese classroom lecture speech contents," in *Proc. of LREC2008*, 2008.

[11] J. Ueberla, "Analysing a simple language model - some general conclusion for language models for speech recognition," *Computer Speech and Language*, Vol. 2, No. 2, 1994, pp. 153–176.