Developing Corpus of Japanese Classroom Lecture Speech Contents

Masatoshi TSUCHIYA[1], **Satoru KOGURE**[2], **Hiromitsu NISHIZAKI**[3], **Kengo OHTA**[4], **Seiichi NAKAGAWA**[4]

[1]Information and Media Center / [4]Department of Information and Computer Sciences,
Toyohashi University of Technology, Japan
[2]Faculty of Informatics, Shizuoka University, Japan
[3]Department of Research Interdisciplinary Graduate School of Medicine and Engineering,
University of Yamanashi, Japan
tsuchiya@imc.tut.ac.jp, kogure@inf.shizuoka.ac.jp,
hnishi@yamanashi.ac.jp, {kohta,nakagawa}@slp.ics.tut.ac.jp

Abstract

This paper explains our developing *Corpus of Japanese classroom Lecture speech Contents* (henceforth, denoted as CJLC). Increasing e-Learning contents demand a sophisticated interactive browsing system for themselves, however, existing tools do not satisfy such a requirement. Many researches including large vocabulary continuous speech recognition and extraction of important sentences against lecture contents are necessary in order to realize the above system. CJLC is designed as their fundamental basis, and consists of speech, transcriptions, and slides that were collected in real university classroom lectures. This paper also explains the difference about disfluency acts between classroom lectures and academic presentations.

1. Introduction

Recently, there is increasing interest in interactive e-Learning systems like exCampus[1], IT's class[2] and Blackboard,[3] because they enable students to learn anywhere and anywhen they want. All of these systems, however, share a big fault: they can treat texts of slides, but not speech. It means that users can not search slides with keywords which are uttered in those speech, although they can search slides with keywords which occur in titles and texts of slides. In order to realize an interactive e-Learning system which can treat both texts and speech, several technologies like spoken document retrieval(Nishizaki and Nakagawa, 2002), video content analysis(Li and Dorai, 2006) and automatic speech summarisation(Hori et al., 2003)(Togashi et al., 2006) are necessary. Especially, robust speech recognition of lecture speech is the most important technology among them.

There are, however, various problems on recognising real classroom lecture speech: speaking styles of teachers, influence of microphones used when recording their speech, noise and/or reverberation of classrooms and language models which cover lecture-related contents. A corpus of classroom lecture speech which is designed particularly for these problem is obviously required, in order to cope these problems.

We already have several corpora of classroom lecture speech. The MIT research group(Park et al., 2005; Glass et al., 2004) has created the corpus, including more 300 hours of English classroom lectures from eight different courses and 80 seminars given on a variety of topics in MIT. This corpus is, however, insufficient to evaluate influence of a generally used lapel microphone, because these data were recorded with an omni-directional microphone under a general classroom environment. LECTRA(L.Lamel et al., 2005; Trancoso et al., 2006), which is the national project in Portugal, includes total 23 Portuguese lectures (approximately 5.2 hours and 44k words included) from two different courses recorded with a lapel and head-mounted microphones. This project also reported the performance of recognising lecture speech and analysed recognition errors. Corpora of general spontaneous speech are possible resources to resolve the described problems. The Rich Transcription (RT) evaluation series [4] that have been started since 2002, are implemented to promote and gauge advances in the state-of-the-art in several automatic speech recognition technologies by using spontaneous speech(Fúgen et al., 2006b)(Fúgen et al., 2006a). In the recent RT evaluation, the tasks of "Speech to Text" (STT), "Speaker Diarization", and "Speech Activity Detection" (SAD) have been evaluated on the three meeting domains. Corpus of Spontaneous Japanese(Maekawa, 2003) (henceforth, denoted as CSJ) is the biggest corpus of spontaneous Japanese including about 1,000 academic presentations and about 1,600 simulated public speech. As each speech included in CSJ was recorded with a headset microphone, it is impossible to use CSJ for evaluating influences caused by microphone types. Furthermore, a corpus of classroom lecture speech is still required, because there is the difference about disfluency acts between academic presentation speech and classroom lecture speech as described later in Section 3..

This paper explains our ongoing project called as *Corpus of Japanese classroom Lecture speech Contents* (CJLC). CJLC is designed as a fundamental basis for developing technologies of robust speech recognition and advanced processing of e-Learning contents, and consists a lot of Japanese classroom lecture speech recorded at several universities. Furthermore, we are going to release CJLC publically for research usage. We hope that CJLC makes a

[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

[4]http://nist.gov/speech/tests/rt/index.htm

Table 1: Statistics of CJLC

# of speakers	15
# of courses	26
# of lectures	89
Duration	3,780 min.

breakthrough in the technologies of spoken language processing for e-Learning contents.

Reminder of this paper is organised as follows: Section 2. describes the detail of CJLC, and Section 3. compares CJLC and CSJ, that is, classroom lecture speech and academic presentation speech. And we conclude in Section 4..

2. Specification of CJLC

As described before, there are several problems on recognising real classroom lecture speech. CJLC is especially designed to resolve two problems among them. The first one is to evaluate influences caused by microphone types under noise and reverberation environment of real classrooms. Speech of CJLC, therefore, are recorded in real classrooms with several microphones, in order to tackle it. The second one is various speaking styles and widespread lecture topics. CJLC covers many speakers at several universities to evaluate influences caused by speaking styles, and consists of many computer science courses, such as physics, electronics, mathematics and information sciences. The later of this section explains the detail of CJLC.

2.1. Structure of CJLC

CJLC is formally defined as a set of classroom lecture data, and each data consists of following items:

- a lecture speech recorded with several microphones,
- its synchronised transcription,
- a presentation slide data (optional, Microsoft PowerPoint formed),
- a timetable of slide show (optional), and
- a list of important utterances (optional).

A lecture speech data and its synchronised transcriptions are provided for all lectures, but a presentation slide data, a timetable of slide show and a list of important utterances are attached to not all lectures. EZ presentator which is an e-Learning software made by Hitachi Advanced Digital Inc. is used to record a timetable of slide show.

Table 1 shows the statistics of CJLC. Because each speaker talks one or more courses, the number of speakers is less than the number of courses. Furthermore, several lectures are recorded for each course as shows in Table 1. 6 lectures among of total 89 lectures contains lists of important utterances, which are annotated by 6 professional researchers.

Table 2: Recording Conditions of CJLC

Microphone type	Recording hardware	Format of speech
wireless lapel (TOA WM-1300)	DAT recorder (Sony TCD-D8)	48KHz 16bit PCM
wired hand held (Sony C-355)		
wired headset (Shure SM10A)	IC recorder (Marantz PMD-671)	16kHz 16bit, PCM

:
0147: でこちらは (F えーと) 発展課題，ですね (F えーと)
0148: やりたい方，(F えーと)
0149: 興味がある (D の) 方はやってみてください
0150: (F えーと) さっきのところって言うのは，データの性質に関わらず，(A エヌ; N) の値のみで決まるデータがどこかというのを
0151: やりました
:
(translated into English)
0147: And here (F well) is the extended exercises, (F well).
0148: The person who want to exercise it, (F well).
0149: Please try (D a) it if you are interested.
0150: (F Well) what I said a little while ago is where is the data decided only by the value of (A enu;N) without the characteristic of data...
0151: I've taught it.
:

Figure 1: An example of transcription

2.2. Recording Condition of CJLC

A lapel microphone is widely used instead of a hand held microphone when recording lectures, because it makes teachers be hands free and does not prevent a lecture, although it drops a recognition performance generally. This means that it is important to investigate performance difference among microphone types on speech recognition and to find a compensation method. A speech data of CJLC, therefore, contains multi-channel data recorded with both a lapel microphone and an other type microphone, unlike previous corpora. For example, a speech data of CJLC contains the data recorded with a lapel microphone and the data recorded with a headset microphone. Table 2 shows various recording conditions for recording speech of CJLC. And more, the speech were recorded at classrooms without special audio equipment, in order to make a corpus under noise and reverberation environment of real classrooms.

2.3. Transcription Format of CJLC

To provide data for acoustic and language model training, we created manual transcriptions of the lecture speech. Each speech was automatically segmented into utterances using the power information of speech described in (Kitaoka et al., 2006; Otsu, 1979). The annotators were instructed to pay careful attention to generating a transcrip-

Table 3: Difference between CJLC and CSJ

	CJLC	CSJ
main target	classroom lectures	lectures in a academic meeting
# of lectures	89	3300
microphone	various microphones (described before)	headset (CROWN CM-311A)
duration per a lecture	long	short
annotation (tagging)	11 tags (CSJ sub-set)	CSJ tags
transcriptions of speech	YES	YES
Slide data	YES (partially)	NO

tion of what was spoken. They were also instructed to annotate speech phenomenas in utterances with the following 11 kinds of tags:

- **(F)** filled pauses,
- **(D)** fragment of content words,
- **(D2)** fragment of functional words,
- **(A)** numerical and alphabetical representation,
- **(W)** corruption,
- **(L)** spoken word(s) with a laugh,
- **(T)** blubbered spoken word(s),
- **(C)** spoken word(s) with cough,
- <**C**> a sound of cough,
- <**B**> a sound of breath,
- <**N**> a noise, and
- <**V**> bubble of voices

These tags are a sub-set of CSJ tag set described in (Koiso et al., 2003), and are compatible to CSJ because CSJ tagging policy is employed when annotating these tags. Tag **(F)**, Tag **(D)** and Tag **(D2)** are especially important to investigate disfluency acts in lecture speech.

Fig. 1 shows an example of transcription of CJLC. Each line, which is corresponding to an utterance unit, consists of two columns: the first column denotes the utterance sequential number in the whole lecture, and the second column shows the transcription of the utterance. In Fig. 1, the Japanese word "えーと", which means "well" in English, is annotated by Tag **(F)** as a filled pause. Tag **(D)** is also employed to annotate the word fragment "の" as a disfluency act.

3. Comparison of CJLC and CSJ

This section explains the difference between classrooms lecture speech and presentation speech.

CSJ is the biggest corpus of spontaneous Japanese, and contains four categories of speech: *academic presentation speech* (APS), *simulated public speech* (SPS), dialogue, and reading. APS is the live recording of academic presentation in 9 different academic societies covering the fields of engineering, social science, and humanities. SPS, on the other hand, is studio recording of layman speaker's speech of about 10–12 minutes, on everyday topics like 'the most delightful/saddest memory of my life'. Table 3 summarises the differences between CJLC and CSJ, which are target speech, transcription tags of phenomenon in spontaneous speech, and slide data.

Table 4: Statistics of CJLC and CSJ

		CJLC	CSJ		
		lectures	APS	SPS	dialogue
# of lectures		89	987	1715	58
# of words / lec.		6636	3358	2122	2613
duration / lec. [sec]		2610	1003	694	765
total duration [hours]		63.0	275.0	330.6	12.3
# of tags / lec.	F	410.3	229.2	118.8	322.2
	D	49.9	44.5	26.0	43.9
	D2	3.9	3.4	1.4	1.4

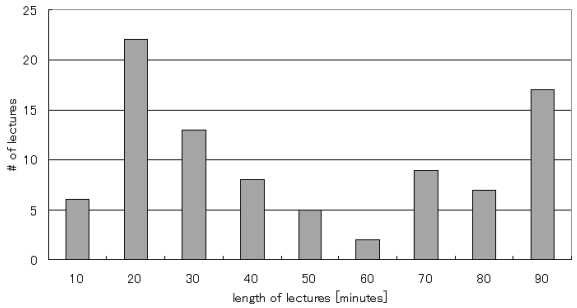

Figure 2: Number of Lectures and Durations

We have compared the phenomenon in spontaneous speech included in CJLC with the one of CSJ. We especially analysed the frequency of filled pauses and disfluently spoken words in each corpus. Table 4 represents the detailed statistics of two corpora. The numbers in Table 4 mean the average values per a lecture (or a dialogue) for each item except the number of lectures and the total duration. Although Table 4 shows that CJLC lectures are longer than CSJ presentations in most cases, it is notable that CJLC lectures can be classified into two categories as shown in Fig. 2. In the latter discussion, the lectures which are shorter than 60 minutes are called as the short lectures, and longer ones are called as the long lectures. Almost the short lectures are guidances to explain practice procedures or exercises. On the other hand, the long lectures are generic classroom lec-

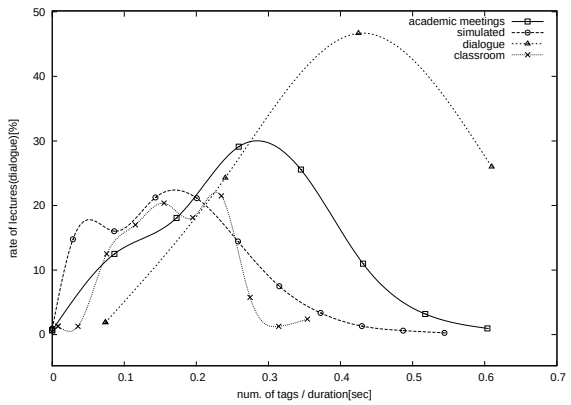

Figure 3: Distribution of Tag F of CJLC and CSJ

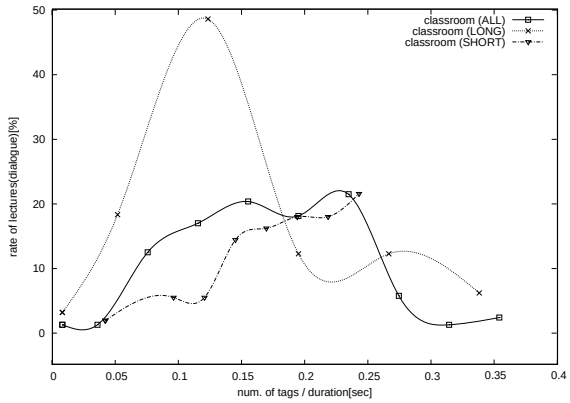

Figure 4: Distribution of Tag F of CJLC Short/Long Lectures

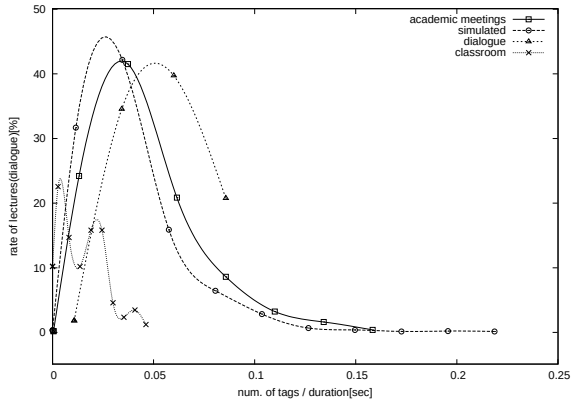

Figure 5: Distribution of Tag D of CJLC and CSJ

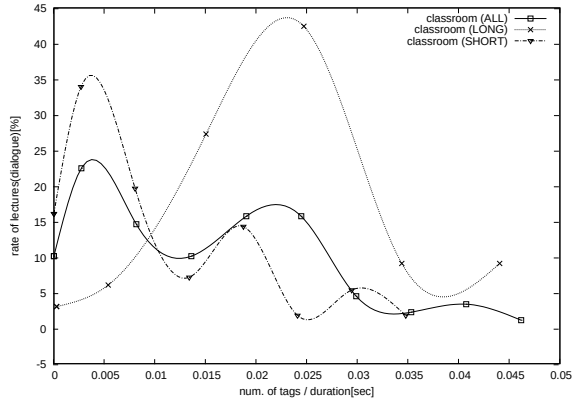

Figure 6: Distribution of Tag D of CJLC Short/Long Lectures

tures.

Fig. 3 shows frequency distributions of filled pauses annotated by Tag F at classroom lectures, APS, SPS and dialogues. As shown in Fig. 3, more filled pauses occur in the dialogue speech of CSJ than in the other types of speech. Classroom lectures of CJLC, APS and SPS share similar frequency distributions of filled pauses. Although Fig. 4 suggests that filled pauses occur more frequently in short lectures than in long ones, we think that there is no serious distinction among them, because both the distribution of filled pauses on short lectures and the one on long ones are still similar to the one on SPS as shown in Fig. 3.

Fig. 5 and Fig. 7 indicate that the number of word fragments in the classroom lectures is less than the number of the speech in CSJ. We think that this is the one of characteristic of CJLC, mainly caused by the fact that speakers of classroom lectures can talk more deliberately than ones of presentations. Fig. 6 shows the frequency distributions of content word fragments annotated by Tag D, and suggests that more disfluency acts on content words occur in long lectures than in short ones. By contrast, Fig. 8 shows that short lectures and long ones are quite similar from the point of view of the average numbers of functional word fragments annotated by Tag D2.

4. Conclusions

This paper explains our developing corpus called as CJLC. Its main aim is to study the current state of classroom lecture speech recognition that is one of the fundamental technologies needed to process the lecture contents. It consists of many real classroom lectures collected at a couple of universities that cover various lecture topics related to information sciences and cover various speaker types. And more, it is possible to evaluate influences caused by various microphones because they contain multi channel recorded speech.

Furthermore, we compared the classroom lectures of CJLC, the presentations and the dialogues of CSJ. The result of analysis represented that the average number of filled pauses included in an utterance is almost the same in the classroom lectures and ones of CSJ. On the other hand, the average number of word fragments caused by disfluency acts in the classroom lectures is less than the number of ones in CSJ speech. We think that this is the one of characteristic of CJLC, mainly caused by the fact that speakers of classroom lectures can talk more deliberately than ones of presentations.

The monitor version of CJLC is already available. Please see `http://www.slp.ics.tut.ac.jp/CJLC/`.

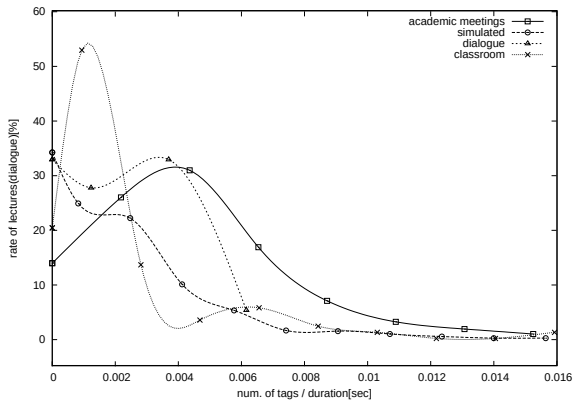

Figure 7: Distribution of Tag D2 of CJLC and CSJ

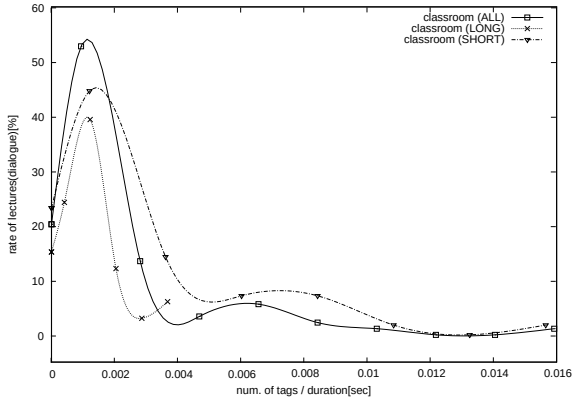

Figure 8: Distribution of Tag D2 of CJLC Short/Long Lectures

We are going to release the formal version of CJLC database to the public limited to only research usage in near future.

5. Acknowledgement

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

6. References

C. Fúgen, M.Kolss, D. Bernreuther, M. Paulik, S. Stúker, S. Vogel, and A. Waibel. 2006a. Open domain speech recognition & translation: Lectures and speeches. In *Proc. of 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2006)*, pages 569–572.

Christian Fúgen, Matthias Wólfel, John W.McDonough, Shajith Ikbal, Florian Kraft, Kornel Laskowski, Mari Ostendorf, Sebastian Stúker, and Kenichi Kumatani. 2006b. Advances in lecture recognition: The isl rt-06s evaluation system. In *Proc. of the 9th International Conference on Spoken Language Processing (Interspeech2006-ICSLP)*, pages 1229–1232.

James Glass, Timothy J. Hazen, Lee Hetherington, and Chao Wang. 2004. 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*, pages 9–12.

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

N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M.Nakayama, Y. Denda, M. Fujimoto, K. Yamamoto, T. Takiguchi, S.Kuroiwa, K. Takeda, and S. Nakamura. 2006. CENSREC-1-C: Development of evaluation framework for voice activity detection under noisy environment. In *IPSJ technical report, Spoken Language Processing (SIG-SLP), Vol.2006, No.107*, pages 1–6.

H. Koiso, Y. Mabuchi, K. Nishizawa, M. Saito, and K. Maekawa. 2003. The specifications of transcriptions version 1.0. In *the Document of Corpus of Spontaneous Japanese*, pages –.

Y. Li and C. Dorai. 2006. Instructional video content analysis using audio information. *IEEE Trans. Audio, Speech, and Language Process.*, 14(6):2264–2274.

L.Lamel, E.Bilinski G. Adda, and J.L. Gauvain. 2005. Transcribing lectures and seminars. In *Proc. of the 9th European Conference on Speech Communication and Technology(EUROSPEECH2005)*, pages 1657–1660.

Kikuo Maekawa. 2003. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, pages 7–12.

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

N. Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-9(1):62–66.

A. Park, Timoty J. Hazen, and James Glass. 2005. 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)*, pages 497–500.

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

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