interspeech 2008 incorporating SST 08

Multi-Modal Recording, Analysis and Indexing of Poster Sessions

Tatsuya Kawahara[*], *Hisao Setoguchi*[*], *Katsuya Takanashi*[*], *Kentaro Ishizuka*[†*], *Shoko Araki*[†]

[*]School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan
[†]NTT Communication Science Laboratories,
Keihanna Science City, Kyoto 619-0237, Japan

Abstract

A new project on multi-modal analysis of poster sessions is introduced. We have designed an environment dedicated to recording of poster conversations using multiple sensors, and collected a number of sessions, to which a variety of multi-modal information is annotated, including utterance units for individual speakers, backchannels, nodding, gazing, and pointing. Automatic speaker diarization, that is a combination of speech activity detection and speaker identification, is conducted using a set of distant microphones, and a reasonable performance is obtained. Then, we investigate automatic classification of conversation segments into two modes: presentation mode and question-answer mode. Preliminary experiments show that multi-modal features on non-verbal behaviors play a significant role in the indexing of this kind of conversations.

Index Terms: multi-modal corpus, poster conversation, speaker diarization, non-verbal information

10.21437/Interspeech.2008-451

1. Introduction

As digital archiving of lectures and meetings has become pervasive, not a few projects on speech technologies oriented for these kinds of audio archives have been conducted. In early 2000s, we compiled the Corpus of Spontaneous Japanese (CSJ)[1], which contains a thousand of academic presentations at technical conferences, recorded using a close-talking microphone. Using this corpus, we have conducted studies on automatic speech recognition (ASR)[2], sentence unit dection and speech summarization[3]. Recordings of oral presentations and seminars were also conducted in European projects such as TED corpus[4] and CHIL projects. Classroom lectures at universities are also being digitally archived, and their automatic transcription and indexing have been also studied under the iCampus project[5][6].

Another target in this direction is a meeting or multiple-party conversation. Projects on meeting archives were initiated by NIST[7] and several European-funded projects such as AMI[8] and CHIL. Since meetings involve multiple participants, it is necessary to determine "who spoke when". This task is referred to as speaker diarization. ASR[9] and dialogue act tagging[10] are also being extensively studied.

In this paper, our new project on multi-modal recording and analysis of poster sessions is introduced. Poster sessions became a norm in many technical conferences, exhibitions, and open laboratories, since they provide more "interactive" characteristics in presentations. Typically, a presenter explains his work to a small audience using a poster, and the audience gives feedback in real time by nodding or backchannels, and occasionally makes questions and comments.

Apparently, the poster session has a mixture of characteristics of lectures and meetings. There are distinct roles in participants, however, anyone can take an initiative in the conversation at a certain point. We expect that this feature provides a new aspect of the research on the analysis of speech communication. Another characteristic of the poster session is that all participants are usually standing, having more freedom in moving heads and bodies, while the existence of the poster makes the participants focus on and point to it. This feature would weigh the importance of annotation and analysis of multi-modal information.

In this paper, we describe the design of the recording environment in Section 2 and the corpus annotation in Section 3. Then, preliminary experiments are reported on speaker diarization and conversation segment tagging using multi-modal information in Section 4 and 5, respectively.

2. Recording Environment: IMADE Room

We are developing a recording environment called the IMADE room at the Faculty of Engineering of Kyoto University, under a project funded by the Japanese MEXT Grant-in-Aid for Scientific Research on Priority Areas: "Info-plosion IT Research Platform". This environment is designed to record audio/visual, human-motion, and physiological data of various kinds of multi-modal human interaction.

Specifically in this work, we used sensor devices to

Accepted after peer review of full paper

Figure 1: Outlook of poster board

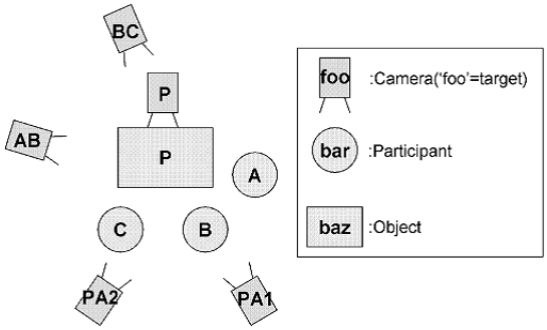

Figure 2: Setting of cameras and poster board

record audio, video, human motion, and eye movements. We used a wireless head-worn microphone (Shure WH-30 XLR) in order to record every participant's voice separately while enabling him to move freely in the room. In addition, we installed an array of eight omni-directional microphones (SONY ECM-77B). They were mounted on the top frame of the poster stand shown in Figure 1.

The IMADE room is equipped with eight built-in cameras for the visual data recording. It is indispensable that significant behaviors of all participants during the session be recorded with at least one camera. By considering these constraints, we carefully designed the setting of poster sessions to be recorded; we assume a session of one presenter using a poster panel and an audience of two persons, and arranged five cameras and the poster board as shown in Figure 2. Here, "P", "A", "B" and "C" indicates the poster, the presenter, and the two persons of the audience, respectively. Note that the poster stand and panel will not occlude the participants B and C from being captured by the camera BC, by arranging the poster mount angled at 22° from the horizontal attitude, as shown in Figure 1.

We also used a motion-capturing system and eye-tracking recorders to record accurate motions and gazing information of the participants, but these have not yet been used in this work.

3. Corpus Collection and Annotation

We have recorded eleven poster sessions so far. We had five different presenters. Each of them had prepared a different poster on his own academic research and conducted two or three sessions. The poster had one main theme and was divided into four sub-topics, which were arrayed in quarters on its surface. The audience in each session had never heard the presentation before. The duration of each session was around 20 minutes.

All speech data were segmented into IPUs (Inter-Pausal Unit) with time labels and transcribed according to the guideline of the CSJ. Annotation of clause boundaries and backchannels were also manually done.

As for annotation of non-verbal information, we focused on gazing, pointing and nodding behaviors. By considering the characteristics of poster conversations, we limited the gazing target to one of other participants and the poster, and the pointing target to the poster. The starting and end time-points were manually labeled for each event based on the video information captured by the cameras. This manual annotation turned out very costly and have been completed for four sessions at this moment. Thus, the following experiments were conducted using the four sessions.

4. Speaker Diarization using Multiple Distant Microphones

In this section, we report speaker diarization experiments using the distant microphones that were installed on the poster stand. The distance from the microphones to the presenter (A) is about 80cm, and the distance to the audience (B and C) is about 130cm. The average distance between the two persons of the audience (B and C) is about 80cm although they can freely move. Note that these kinds of information (the number and location of the speakers) are not used in the speaker diarization experiments. The signal-to-noise ratio (SNR) of the captured audio was 0 to 6.5dB because of the environmental noises such those from computers in the room.

The process of speaker diarization consists of speech activity detection, which detects the speech segments uttered by either participants from the audio signals, and speaker indexing, which classifies the detected speech segments into one of the speakers.

4.1. Speech Activity Detection

In order to cope with environmental noises including burst noise, we proposed a voice activity detector that uses the power ratios of periodic to aperiodic component of the observed signals[11]. This method is called

PARADE (periodic to aperiodic component ratio based detection) henceforth. We applied the method to the recorded signals obtained from one microphone of the array.

4.2. Speaker Indexing

Speakers of the detected speech segments are identified by classifying estimated DOA (Direction Of Arrival) of the speech. In this method, the speech segments that came from a certain spatial region are considered as those produced by an identical speaker. The DOA is estimated by utilizing three microphones of the array, which are located in a triangle position. The DOA estimation is performed using the generalized cross correlation method with the phase transform (GCC-PHAT)[12], which is widely-used in the meeting task.

Then, the estimated DOAs are classified into the speakers. In order to classify the DOAs without a priori knowledge about the number of speakers, the leader-follower clustering algorithm[13] is applied. The algorithm conducts online clustering by generating a new centroid when a speech segment is observed from a new spatial region.

4.3. Results

The specification of the evaluation set of the speaker diarization is shown in Table 1. As an evaluation measure, we adopt diarization error rate (DER), which is used in NIST Rich Transcription[7]. The DER accounts for missed speech time (MST), false-alarm speech time (FST), and speaker error time (SET), and is calculated by dividing the sum of FST, SET and MST by the total length of the recorded data. The evaluation criteria also follow that provided by NIST. The margin to tolerate the difference between the system outputs and the correct labels is 250ms.

Table 2 shows the speaker diarization results. The sum of MST and FST rates shows performance of the speech activity detection, and the error rate is only 1.3-6.6%, demonstrating the robustness of the proposed PARADE method against environmental noises.

On the other hand, SET rate is relatively large because the current speaker classification method relies on the DOA estimation. The performance is degraded by speaker movements, for example, the audience approaches the presenter or the poster when he raises questions. Moreover, the DOA of some directional noises such as the computer noise is sometimes estimated as that of the speech signal when the power of speech is lower than the noise in the speech segment. The robustness of speaker indexing against speaker movements and directional noise sources must be improved in the future.

Table 1: Statistics of utterance duration (sec.) in test set

	A	B	C	total
session 1	824	287	32	1037
session 2	789	129	129	913
session 3	1068	59	178	1150
session 4	1175	32	200	1291

Table 2: Speaker diarization results

	MST	FST	SET	DER
session 1	2.8	2.2	27.1	32.0
session 2	3.0	3.6	17.5	24.1
session 3	0.6	3.4	17.9	21.9
session 4	0.7	0.6	17.2	18.5

5. Classification of Conversation Mode

In this section, we address another classification of speech segments; that is to classify segments into presentation mode and question-answer mode. In the presentation mode, the presenter keeps an initiative and mainly gives an explanation on his work, accompanied by some feedback from the audience, such as backchannels and short comments. In the question-answer mode, one of the audience takes an initiative and raises questions, which are replied by the presenter. The annotation was done manually by considering who takes an initiative in the conversation segment.

We presume that this kind of indexing would be useful when browsing a recorded archive of the poster sessions, and investigate an automatic indexing method.

5.1. Approach based on Speaker Diarization

The first approach is to make use of the speaker diarization information. Apparently, when one of the audience raises a question, he should speak for a certain period. And this event will occur between some interval. So, we set thresholds for the duration and the interval of the utterances (IPUs) of the audience, and investigated how accurately we can detect the utterances which belong to the question-answer mode.

Here, we used the correct labels of speaker diarization rather than the results of the previous section to see the upper bound, and experimentally tuned the threshold values, however, the detection accuracy (=correctly detected QA utterances divided by the total) is only 61%. There are a number of question utterances whose duration is shorter than the threshold, while there are many long feedback utterances during the presentation mode. The result suggests the limitation of the simple classification based on speaker diarization without transcribing the utterances.

Table 3: Distribution of non-verbal behaviors

mode	duration	mutual gazing	joint attention	pointing
presentation	568	61 (.11)	295 (.52)	181 (.32)
question -answer	670	225 (.34)	190 (.28)	120 (.18)

(upper row: sec., lower row: ratio)

Table 4: Automatic identification results of conversation mode

mutual gazing	joint attention	pointing	combined (LDA)
68%	69%	63%	72%

5.2. Approach using Multi-Modal Information

Next, we investigate the use of multi-modal information. In the presentation mode, the presenter is often pointing to the poster, thus gazing it. Accordingly, the audience should also be gazing the poster, which suggests that joint attention (to the poster) is dominant during the presentation mode. On the other hand, in the question-answer mode, the presenter and the person who raises the question should make a dialogue, which implies that mutual gazing (between the two persons) is dominant. In this mode, pointing to the poster may be less frequent.

In this experiment, we used the manual annotation of pointing and gazing information as a preliminary evaluation. Table 3 lists distributions of these behaviors for each mode, averaged over the test-set sessions. In each entry, the absolute duration in seconds is given in the upper row, and its ratio in the total duration is in the lower row. The differences between the mean values of the two modes are statistically significant ($p < 0.001$) for all three features; it is confirmed that there are more mutual gazing and less joint attention and pointing in the question-answer mode.

Finally, we conducted automatic identification of the mode using these features. The session is segmented into units of 10 seconds, and the occurrence ratio of each behavior is computed for each segment. We also performed a linear discriminant analysis (LDA) to estimate weights of the features when combining them with a linear function. This is done with the one-leave-out manner. The classification results are summarized in Table 4. It is observed that each feature has discriminant information and the combination of the three features has a synergetic effect.

6. Conclusions and Future Directions

In this paper, we introduced a new project on the multi-modal analysis of poster sessions. The poster session has a mixed characteristic of presentation and discussion. As a preliminary analysis, we conducted automatic identification of the two modes. It is suggested that multi-modal features focusing on non-verbal behaviors provide useful information. As the experiment in the previous section relies on the manual annotation, one of the future works is fully automatic extraction of these features.

The next step is to intergrate the non-verbal information with verbal information. In this paper, we demonstrated that automatic speech activity detection can be reliably performed even with distant microphones. We plan to improve speaker indexing and investigate automatic transcription as well as exploiting prosodic information.

Acknowledgements: The authors are deeply grateful to the members of Prof. Nishida & Sumi's Laboratory and the members of Prof. Nakamura's Laboratory of Kyoto University for joint collaboration in the IMADE room project.

7. References

[1] S.Furui. Recent advances in spontaneous speech recognition and understanding. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 1–6, 2003.

[2] H.Nanjo and T.Kawahara. Language model and speaking rate adaptation for spontaneous presentation speech recognition. *IEEE Trans. Speech & Audio Process.*, 12(4):391–400, 2004.

[3] T.Kawahara, M.Hasegawa, K.Shitaoka, T.Kitade, and H.Nanjo. Automatic indexing of lecture presentations using unsupervised learning of presumed discourse markers. *IEEE Trans. Speech & Audio Process.*, 12(4):409–419, 2004.

[4] E.Leeuwis, M.Federico, and M.Cettolo. Language modeling and transcription of the TED corpus lectures. In *Proc. IEEE-ICASSP*, volume 1, pages 232–235, 2003.

[5] A.Park, T.Hazen, and J.Glass. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In *Proc. IEEE-ICASSP*, volume 1, pages 497–500, 2005.

[6] C.Chelba and A.Acero. Indexing uncertainty for spoken document search. In *Proc. INTERSPEECH*, pages 61–64, 2005.

[7] J.S.Garofol, C.D.Laprun, and J.G.Fiscus. The RT-04 spring meeting recognition evaluation. In *NIST Meeting Recognition Workshop*, 2004.

[8] S.Renals, T.Hain, and H.Bourlard. Recognition and understanding of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop Automatic Speech Recognition & Understanding*, 2007.

[9] J.Fiscus, J.Ajot, and J.S.Garofol. The rich transcription 2006 evaluation overview and speech-to-text results. In *NIST Meeting Recognition Workshop*, 2006.

[10] E.Shriberg, R.Dhillon, S.Bhagat, J.Ang, and H.Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proc. SIGDial*, pages 97–100, 2004.

[11] K.Ishizuka, T.Nakatani, M.Fujimoto, and N.Miyazaki. Noise robust front-end processing with voice activity detection based on periodic to aperiodic component ratio. In *Proc. INTERSPEECH*, pages 230–233, 2007.

[12] C.H.Knapp and G.C.Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust., Speech & Signal Process.*, 24(4):320–327, 1976.

[13] R.O.Duda, P.E.Hart, and D.G.Stork. *Pattern Classification*. John Wiley & Sons, New York, 2000.