Constructing a Multi-Point Video Conference Corpus Labeled by Participants' Feelings for Developing "Video Conference Grasper"

Akira NAKAYAMA, Masamichi HOSODA, Minoru KOBAYASHI and Satoshi IWAKI
nakayama.akira@lab.ntt.co.jp
NTT Cyber Solutions Laboratories, NTT Corporation
1–1 Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239–0847 Japan

Abstract

Our target is a teleconferencing system that allows a busy user to grasp the key points in live and recorded teleconferences. To develop this system, we believe that information beyond the physical attributes of speech such as utterance characteristics must be collected and processed. Our first step, reported here, is to construct a teleconference corpus that contains labels indicating the participants' feelings. A preliminary analysis of the correlation of labeled topic boundaries and the changes in participants' utterances shows the possibility of recognizing a conversation's structure from changes in the amounts of participant utterances.

1. Introduction

Modern life is characterized, in part, by the spread of broadband Internet connections, powerful computers, and social flexibility. It results in a significant increase in the importance and frequency of teleconferences; Internet-based teleconferencing can dramatically lower costs.

In order to identify the potential and the basic problems of conventional Internet-based teleconferencing, we used a commercial Internet video meeting system to connect multiple homes and offices and examined its operation for about one year. This extended trial showed that the average user found it very difficult to attend all meetings in their entirety because of interruptions such as telephone calls, visitors, and regular work. Namely, the meeting system forces the user to connect to the real and cyber communication spaces simultaneously. The users indicated the need for some support tools such as a meeting minutes creation function and a function that allows the user to grasp the outline of a meeting quickly and accurately.

We recognized the importance of these and other similar functions that support participation in the communication space. The key functions are, we believe, those that present to the user only changes in meeting atmosphere, user-specified topics, and important arguments. We cover these functions with the overarching title "Video Conference Grasper." The research areas related to Video Conference Grasper are discussed from the viewpoints of groupware and media processing.

Several papers on groupware describe systems for creating multimedia minutes. Most of these systems require a dedicated person to set the structure and write the utterances down [1]. In the asynchronous and synchronous integrated groupware system called ASSIST, a meeting participant must input data on the intention (support, opposition, proposals, etc.) of each utterance, which means that proceedings cannot be processed in real-time [2]. Although one example showed the technique of creating, in real-time, a digest by using intentional verbal cues such as "applause," the manual effort involved is still excessive and the resulting digest is of limited value [3]. The conventional systems do not allow the user to access the meeting structure by using subtle communication cues such as "disagreement by silence" and "rejection of an argument as shown by body posture."

With regard to processing dialog by computer, some papers have tackled structure atmosphere analysis using non-linguistic information based on large-scale speech dialog corpora as well as the continuous speech recognition of meeting dialogs [4]. Galley et al. discussed the amount of overlapping speech at topic shifts and sudden changes in speaker activities [5] with the ICSI Meeting corpus [6]. Warede et al. identified the characteristics of utterance pitch frequency and power changes which are considered to be key communication inflection points as discerned by the analysis of the same corpus by a third person [7]. Past research could be considered to have assumed that signal processing could extract the structure and atmosphere of conversations. They have achieved some success but none explicitly deals with the interaction between speakers in situations where a chairperson is present. Moreover, there is no corpus whose contents includes tags created by the participants themselves. This information, the tags, is important when trying to identify the atmosphere and the key point of the meetings. Participants' tags can be more accurate than those of a third person for identifying internal information, such as participants' feelings and judgements. Few papers have dealt with video conferences corpus, which usually needed to allow consideration of the augmented communication paths such as text chatting and document share, and speech patterns change due to video conference systems [8].

In order to advance the research in this field, we focused on constructing a corpus that includes tags that identify the users' feelings and their perceived intention in addition to text chatting and documents share operation logs. This paper first describes the features of the multi-point multimedia meeting recording system used for corpus construction and corpus design and then reports a recording experiment and a preliminary statistical analysis of the changes in utterance amounts at topic segment shifts.

2. Meeting Recording System

This section outlines the meeting system design and configuration.

[Requirements] We designed the meeting recording system as a research platform that meets three requirements.

1. Naturalness: provides the quality and space needed to encourage novices to talk naturally.
2. High-quality: The quality of the recorded video and speech are sufficient to pėrmit image and speech analysis.
3. Synchronicity: Multimedia data are stored together with time stamps.

[Hardware Configuration] An outline of the system is shown in Fig. 1. It consists of four client PCs equipped with two displays for facial pictures and shared documents, a USB camera (Logicool QCAM Pro4000), a head set microphone (SONY DR-140), and two server PCs for a text event data inclusion and multi-point audio-visual communication. These PCs are connected via a private gigabit network. Each of the input-and-output devices of the four client PCs is set in a soundproof chamber, a noise-free room so that no direct auditory/visual connection is possible between the 4 participants; external noise is shut out. A 4-channel color multiplexer was used to collect and record the image and sound from the digital camcorders installed in the soundproof chambers. An audio mixer was also installed.

[Software Configuration] We used Flash Communication ServerMX (MacroMedia) which is easy to customize and developed a client program that runs on the Flash player to realize multi-point video meeting and recording. Moreover, the window share function of NetMeeting3.0 and PowerPoint (Microsoft) were combined to realize document sharing. The text chatting and document sharing operations were stored in the HTTP server by using the HTTP POST method. The player, which reproduces a recorded meeting, was also developed by the same method.

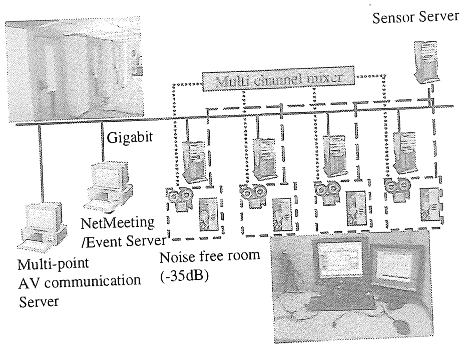


Figure 1. System configuration

Table 1. Overview of system performance

Item	Performance
Face image frame rate	15 fps
Face image frame size	160*120
Audio sampling rate	22050 Hz
Audio delay approximately	220 msec
Simultaneous attendees	4 adults

[Performance] An overview of system performance is shown in Table 1. The speech delay of 220 ms does not affect the naturalness of conversation according to Kurita et al. [9]. Any offset between the image and speech streams was insignificant.

3. Corpus construction

3.1. Aim

The corpus is to provide the raw data from which we can extract useful information (feelings, attitudes, and topic segment shifts) and develop a useful processing algorithm.

For this purpose, we collected speech and image streams and had the subjects themselves tag each "utterance" with feelings, attitudes, and value judgments, as well as getting them to identify and name topic segments; this information was gathered by the questionnaire method.

3.2. Discussion details and questionnaire design

In order to acquire plentiful clues from an animated discussion, the meetings were designed with the goals of developing plans and/or making decisions. The subjects were comfortable with each other (every group consisted of friends). Of the 8 trials (meetings), seven used Japanese adults while one used Chinese adults in an attempt to gather some cross-cultural information. Three agendas were set: creating a name for a new beverage, designing a new electric appliance, and selecting a set of survival tools. They were intended to get the participants talking freely without special knowledge.

The questionnaire listed every utterance made in the meeting to gather as much information as possible. Four categories (consisting of seven items) of information were collected as shown below. These 7 items were selected from the viewpoints of the development of atmosphere, key points and topic transition extraction algorithms.

Feelings: The subject was instructed to indicate first his/her support of the utterance (strong support +2, support +1, neutral 0, antipathy −1, strong antipathy −2) and his/her interest in the utterance (strong stimulation +2, stimulation +1, neutral 0, depression −1, strong depression −2). This scoring follows the 2-dimensional feeling model [10].

Value judgment: The subject was instructed to indicate his/her judgment of the importance of the utterance (very important +2, important +1, neutral 0). He was told that this importance covered items such as the contribution of the utterance to the meeting, the importance of the idea, and the dignity of the utterance. Next, he scored the novelty of the utterance (very novel +2, novel +1, completely expected 0). This covered items such as the impact on the flow of conversation.

Attitude to opinions: An utterance, other than a fact, that was intended to persuade was defined as an opinion. The subject was asked to indicate the strength of each opinion (strong opinion +2, weak opinion +1, neutral 0) and his/her agreement with it (strong agreement +2, agreement +1, neutral 0, disagreement −1, strong disagreement −2).

Topic transition: To investigate how the participants perceived the structure of the meetings, we asked them to set topic segment boundaries and give a appropriate title as they pleased to each segment. The participants were told to set the topic segment boundaries by their intuition.

3.3. Design of the minutes format

To investigate how the subjects grasped the macroscopic features, impressions, and flow of meetings, we asked each of them to create the minutes of the meeting they attended. They were provided with access to the video recording of the meeting and were asked to indicate the general flow of the meeting, main point of the meeting, important utterances, and climax parts by indicating whom, when, and what. They were also asked to describe what was forgotten but necessary to make a minutes by the same format.

3.4. General flow of meeting recording

Each meeting proceeded in the following steps: explanation, practice, recording, minutes creation, making the questionnaire sheet, and filling in the questionnaire. Each trial took about 8 hours.

The participants understood the purpose of the experiment; it was for research and the desire was to record the state of natural meetings in detail. For rehearsal, the participants used the meeting system to play a few games and hold everyday conversations. We showed them the topic and asked them to set the agenda. After being given the agenda, the participants spent some time thinking about their opinions. Prior to starting the meeting, they were told they would be asked to write a minutes after the meeting, and they could freely take notes on the note sheets prepared in the room. They were told to select a chair person first, and then to start the meeting in the language of their native country (Japanese or Chinese).

After the meeting was recorded, transcriptions were generated; we wrote down every utterance in a 10 minute section of the conversation selected arbitrarily and created the questionnaire using the 10 minute section of the meeting selected. The subjects first created the minutes, and then filled in the questionnaire while reviewing the recording of the meeting.

3.5. Rule of transcription

To generate the transcripts of the meetings, we defined the following; an utterance unit was "utterance of one person isolated by more than 400 msec from adjacent utterances of the same person." The start and end times of utterance units were identified at the level of millisecond. Silent sections longer than 100 msec within an utterance unit were identified as pauses and their lengths were noted [11].

3.6. Recording results: basic statistics

The trials are described in Table 2. The normalized number and total duration of utterances per 60 seconds made by each subject are shown in Table 3 for comparison. Bold characters indicate the chairperson's data. It is clear that the chairperson tends to made longer utterances than the other subjects. The differences are significant according to the Wilcoxon test ($p < 0.01$). This is due to the special role played by the chairman and suggests that special analysis of the chairpersons' utterances is important.

Table 2. Overview of recorded data

Trial	Topic	Participants [M: Male, F: Female]	Duration [min]	text chat [times]	Utterance [times]	Language used by Participants
1	survival	F4/telephone operator	29	164	983	Japanese
2	survival	F4/filing clerks	34	32	888	Japanese
3	name	F4/former architecture students	62	54	1603	Japanese
4	name	M2F2/vocational school students	14	35	624	Japanese
5	name	M1F3/band members	44	43	1535	Japanese
6	plan	M4/university students	36	26	1001	Japanese
7	name	M2F2/Chinese	26	0	526	Chinese
8	name	F4/clerks	21	32	653	Japanese

Table 3. Normalized number of utterances, total utterance duration per 60 seconds

		Utterance[Times]/Utterance[sec]			
Trials	Topics	Speaker1	Speaker2	Speaker3	Speaker4
1	survival	**11.6/18.2**	6.02/4.70	6.02/28.8	9.99/13.3
2	survival	8.99/11.0	**10.7/11.6**	8.28/12.8	4.52/5.76
3	name	7.14/8.49	3.53/4.49	**8.26/14.3**	6.84/9.07
4	name	**13.2/18.1**	12.8/14.7	7.78/6.75	8.04/7.98
5	name	**17.8/24.9**	5.46/6.17	8.37/10.3	3.28/4.06
6	plan	8.40/10.3	7.31/8.17	5.43/8.10	**6.89/6.38**
7	name	**12.1/39.1**	1.85/7.05	1.66/1.81	3.50/4.35
8	name	**16.5/24.1**	5.39/9.16	4.20/5.10	4.72/7.95

Table 4. Concurrence of topic segment annotation

Trials	κ	κ(allowance 1 utterance gap)	Topic boundaries
1	0.017	–	–
2	0.077	–	–
3	0.39	0.57	11
4	0.35	0.46	9
5	0.31	0.39	10
6	0.39	0.43	6
7	0.40	0.53	7
8	0.35	0.35	5

4. Correlation of the amount change of utterance and topic segment

4.1. Reliability of annotation of topic boundaries

We examined the characteristics of topic segment boundaries as the first step toward "Conference Grasper." To measure the degree of coincidence among the subjects' decision of topic segment boundaries (TBs), we calculated Cohen's κ[12]. A typical calculation result is shown in Table 4.

In many cases, κ values from 0.4 to 0.6 were obtained. This range indicates fair coincidence [12]. The number of the segment boundaries that were commonly identified by at least three subjects with a gap of not more than 1 utterance is also shown. The result indicates that the subjects' intuitions as to the topic segment boundaries showed fair coincidence.

4.2. The length of topic boundaries

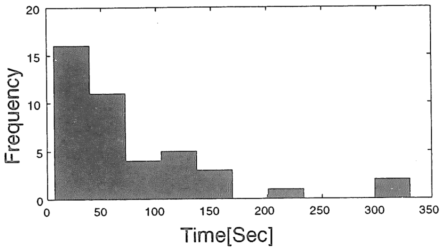


Figure 2. The length of topic boundaries

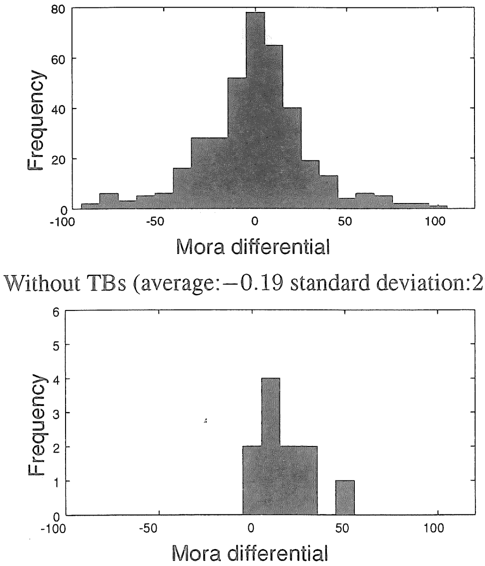
A histogram of segment length is shown in Fig. 2. The start time of a segment is the start time of the utterance unit of the beginning just behind a TB, and end time of a segment is the end time of the utterance unit in front of a next topic segment boundary. The average length was 75 seconds and 50% of the segments ended within 50 seconds; 90 percent ended within 150 seconds. These values are short compared to prior research (an average of about 540 seconds [5]), due to differences in the experimental technique and the task set, but we note that the distribution trend is similar. This fact may reflect the recursive structure of topic segment boundaries.

4.3. The number of utterance mora

The correlation of the number of utterance mora and segment boundaries was analyzed. In order to understand the utterance activities before and after segment boundaries, we investigated the changes in the number of mora before and after every utterance unit in the 3rd meeting. A histogram of the difference in the number of mora is shown in Fig. 3(a). The histogram of the difference in the number of mora with segment boundaries following the same method is shown in Fig. 3(b).

There is a remarkable difference between these two distributions. Fig. 3(b) suggests that there is a significant increase of the utterance amount at the topic segment boundaries: the utterance amount decreases just before segment boundaries, and then increases just after the segments. Wilcoxon's test shows that this tendency is significant ($p < 0.01$). On the other hand, there is no prominent change in the cases that show no topic segment boundaries (Fig. 3(a)). These results indicate the tendency that we speak less just before the topic boundaries, and speak a lot just after the boundaries. This tendency gives a useful clue to automatic understanding of the conversation structures.

Among the 11 long utterances just after the topic segment boundaries, 7 utterances were made by the chairperson. This suggests that the chairpersons' long utterances may also give a useful clue to understanding the conversation structure.



(a) Without TBs (average:-0.19 standard deviation:29.1)

(b) With TBs (average:17.5 standard deviation:14.3)

Figure 3. Histograms of the difference in the number of mora with and without topic boundaries

5. Conclusion

The acquisition of an annotated video meeting corpus was reported, and the possibility of recognizing a conversation's structure from changes in the participants' volume of utterances was shown as was the special role played by the chairperson; there is a tendency that the number of utterances significantly increases just after a topic boundary. We will make a more detailed analysis and to strive to establish the tools needed to automatically extract the structure and atmosphere of meetings.

References

[1] Haruhiko Kaiya, et al., "Preliminary Experiments of A Computer System for Face-to-face Meetings," Systems and Computers in Japan, Vol.28, No.2, pp. 21–32, 1997.

[2] Michiru Tanaka and Yoshimi Teshigawara, "Design of a Development Environment for Web-based Asynchronous and Synchronous Integrated Groupware Systems," Proceedings of IEEE 23rd International Conference on Distributed Computing Systems Workshops, pp.582–587, 2003.

[3] Akihito Kawaguchi, et al., "A Digest Making Method for Helping Users to Join to Teleconference from Halfway," Journal of Information Processing Society of Japan, Vol.42, No.12, pp.3031–3040, 2001.

[4] Alex Waibel, et al.,"Advances in Automatic Meeting Record Creation and Access," Proceedings of ICASSP2001, vol.1, pp.597–600, 2001.

[5] Michel Galley, et al., "Discourse Segmentation of Multi-Party Conversation," Proceedings of 41st Annual Meeting of ACL, pp.562–569, 2003. `http://www1.cs.columbia.edu/~galley/`

[6] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," In Proc. of ICASSP-03, 2003.

[7] Britta Wrede and Elizabeth Shriberg, "Spotting Hot Spots in Meetings: Human Judgments and Prosodic Cues," Proceedings of EUROSPEECH 2003, pp.2805–2808, 2003.

[8] Abigail J. Sellen, "Speech Patterns in Video-Mediated Conversations," Proceedings of CHI'92, pp.49–59, 1992.

[9] T. Kurita, S. Iai, N. Kitawaki, "Assessing the Effects of Transmission Delay —Interaction of Speech and Video—," Proceedings of 14th Int. Symposium on Human Factors in Telecommunications, HFT'93, pp. 111-120, 1993.

[10] Byron Reeves and Clifford Nass, "The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places," Univ. of Chicago Pr, 1998.

[11] Akira Ichikawa, et al.,"Standardising Annotation Schemes for Japanese Discourse," Proc 1st Int'l Conf. on Language Resource and Evaluation, pp.731–736, 1998.

[12] Fleiss, J. L.,"Measuring Nominal Scale Agreement Among Many Raters," Psychological Bulletin, Vol.76, pp.378–382, 1971.