# Chapter 9
# Smart Posterboard: Multi-modal Sensing and Analysis of Poster Conversations

**Tatsuya Kawahara**

**Abstract** Conversations in poster sessions in academic events, referred to as poster conversations, pose interesting and challenging topics on multi-modal multi-party interactions. This article gives an overview of our CREST project on the smart posterboard for multi-modal conversation analysis. The smart posterboard has multiple sensing devices to record poster conversations, so we can review who came to the poster and what kind of questions or comments he/she made. The conversation analysis combines speech and image processing such as face and eye-gaze tracking, speech enhancement and speaker diarization. It is shown that eye-gaze information is useful for predicting turn-taking and also improving speaker diarization. Moreover, high-level indexing of interest and comprehension level of the audience is explored based on the multi-modal behaviors during the conversation. This is realized by predicting the audience's speech acts such as questions and reactive tokens.

**Keywords** Multi-modal · Conversation analysis · Speech processing · Posterboard

## 9.1 Introduction

Speech and image processing technologies have been improved so much that their target now includes natural human-human behaviors, which are made without being aware of interface devices. Examples of this kind of direction include meeting capturing [1] and conversation analysis [2]. We have conducted the CREST project, which focused on conversations in poster sessions, hereafter referred to as poster conversations [3, 4]. Poster sessions have become a norm in many academic conventions and open laboratories because of the flexible and interactive characteristics. In most cases, however, paper posters are still used even in the ICT areas. In some cases, digital devices such as LCD and PC projectors are used, but they do not have sensing devices. Currently, many lectures in academic events are recorded and distributed via Internet, but recording of poster sessions is never done or even tried.

T. Kawahara (✉)
Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
e-mail: kawahara@i.kyoto-u.ac.jp

T. Nishida (ed.), *Human-Harmonized Information Technology, Volume 1*,
DOI 10.1007/978-4-431-55867-5_9

Poster conversations have a mixture characteristics of lectures and meetings; typically a presenter explains his/her work to a small audience using a poster, and the audience gives feedbacks in real time by nodding and verbal backchannels, and occasionally makes questions and comments. Conversations are interactive and also multi-modal because participants are standing and moving unlike in meetings. Another good point of poster conversations is that we can easily make a setting for data collection which is controlled in terms of familiarity with topics and other participants and yet is "natural and real".

The goal of the project is signal-level sensing and high-level analysis of human interactions. Specific tasks include face detection, eye-gaze detection, speech separation, and speaker diarization. These will realize a new indexing scheme of poster session archives. For example, after a long session of poster presentation, we often want to get a short review of the question-answers and feedbacks from the audience.

As opposed to the conventional "content-based" indexing approach which focuses on the presenter's speech by conducting speech recognition and natural language analysis, we adopt an "interaction-oriented" approach which looks into the audience's reaction. Specifically we focus on non-linguistic behaviors such as backchannel, nodding and eye-gaze information, because the audience better understands the key points of the presentation than the current machines. An overview of the proposed scheme is depicted in Fig. 9.1.

We have designed and implemented a research platform for multi-modal sensing and analysis of poster conversations. From the audio channel, utterances as well as laughter and backchannels are detected. Eye-gaze and nodding are also detected by using video and motion sensing devices. Special devices such as a motion-capturing system and eye-tracking recorders are used to make ground-truth annotation, but only video cameras and distant microphones are used in the practical system.

We also investigate high-level indexing of which segment was attractive and/or difficult for the audience to follow. This will be useful in speech archives because people would be interested in listening to the points other people liked. However, estimation of the interest and comprehension level is apparently difficult and largely subjective. Therefore, we turn to speech acts which are observable and presumably related with these mental states. One is prominent reactive tokens signaled by the

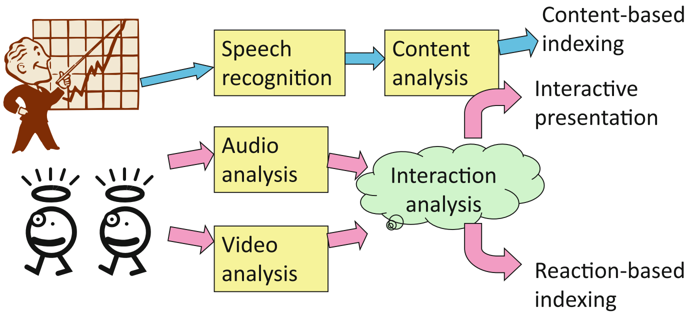

**Fig. 9.1** Overview of multi-modal interaction analysis

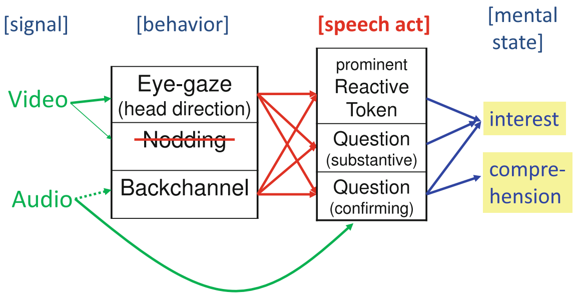

**Fig. 9.2** Proposed scheme of multi-modal sensing and analysis

audience and the other is questions raised by them. Prediction of these speech acts from multi-modal behaviors is expected to approximate the estimation of the interest and comprehension level. The scheme is depicted in Fig. 9.2.

## 9.2 Overview of System and Corpus

### *9.2.1 Smart Posterboard System*

We have designed and implemented a smart posterboard, which can record a poster session and sense human behaviors. Since it is not practical to ask every participant to wear special devices such as a head-set microphone and an eye-tracking recorder and also to set up any devices attached to a room, all sensing devices are attached to the posterboard, which is actually a 65-in. LCD screen. Specifically, the digital posterboard is equipped with a 19-channel microphone array on the top, and attached with six cameras and two Kinect sensors. An outlook of the smart posterboard is given in Fig. 9.3. A more lightweight and portable system is realized by only using the Kinect sensors, which captures audio and video signals.

### *9.2.2 Multi-modal Corpus of Poster Conversations*

We have recorded a number of poster conversations for multi-modal interaction analysis [3, 5]. In each session, one presenter (labeled as "A") prepared a poster on his/her own academic research, and there was an audience of two persons (labeled as "B" and "C"), standing in front of the poster and listening to the presentation. Each poster was designed to introduce research topics of the presenter to researchers or students in other fields. The audience subjects were not familiar with the presenter and had not heard the presentation before. The duration of each session was 20–30 min. Some presenters made a presentation in two sessions, but to a different audience.

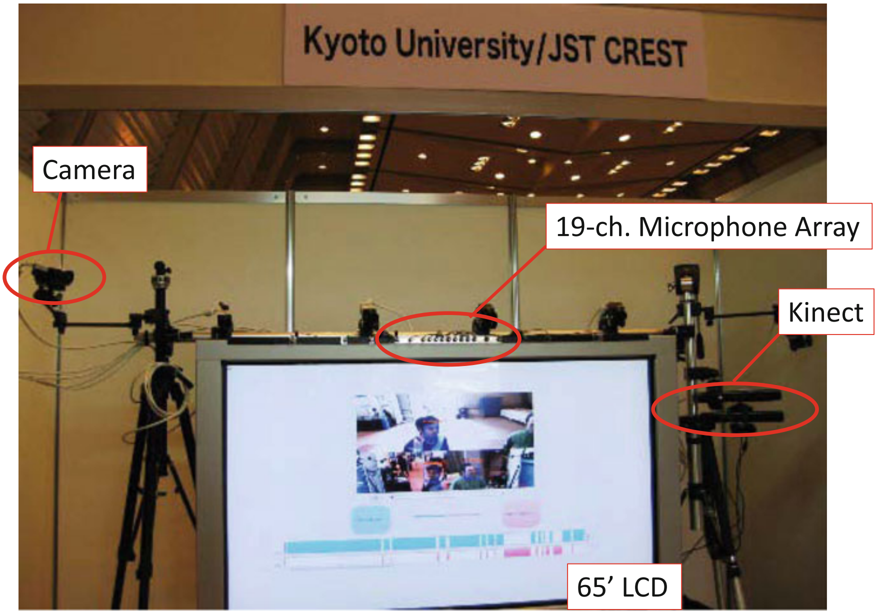

**Fig. 9.3** Outlook of smart posterboard

All speech data were segmented into IPUs (Inter-Pausal Unit) and sentence units with time and speaker labels, and transcribed according to the guideline of the Corpus of Spontaneous Japanese (CSJ) [6]. Fillers, laughter and verbal backchannels were also manually annotated. While fillers are usually followed by utterances by the same speaker, backchannels are uttered by themselves.

For the ground-truth annotation, special multi-modal sensing devices such as a motion capturing system were used while every participant wore a wireless headset microphone and an eye-tracking recorder or a magnetometric sensor. In the early phase of the project, eye-gaze information was derived from the eye-tracking recorder and the motion capturing system by matching the gaze vector against the position of the other participants and the poster. But their calibration and post-processing are very time-consuming. In the latter phase of the project, the magnetometric sensor were adopted to estimate head orientations instead of precise eye-gaze.

### *9.2.3 Detection of Participants’ Eye-Gaze and Speech*

Detection of participants and their multi-modal feedback behaviors such as eye-gaze and speech using the smart posterboard (green lines in Fig. 9.2) is explained. It is realized with multi-modal information processing, as shown in Fig. 9.4, and briefly explained in the following subsections.

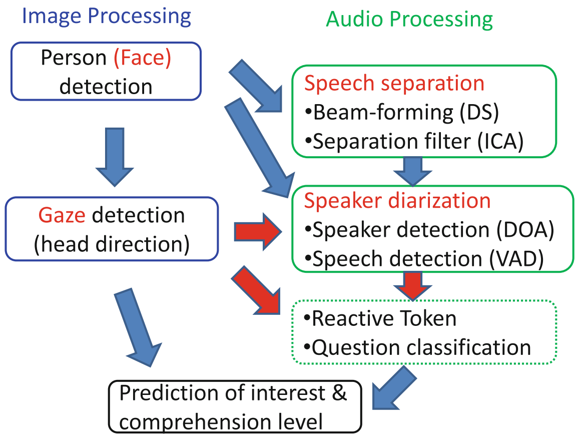

**Fig. 9.4** Process flow of multi-modal sensing

#### 9.2.3.1 Face and Eye-Gaze Detection

Kinect sensors are used to detect the participants' face and their eye-gaze. As it is difficult to detect the eye-ball with the Kinect's resolution, the eye-gaze is approximated with the head orientation. A preliminary analysis using the eye-tracking recorder showed that the difference between the actual eye-gaze and the head orientation is 10° on average, but it is much smaller when the participants look at the poster. The process of the face and the head orientation detection is as follows [7]:

1. Face detection
   Haar-like features are extracted from the color and ToF (Time-of-Flight) images to detect the face of the participants. Multiple persons can be detected simultaneously even if they move around.
2. Head model estimation
   For each detected participant, a three-dimensional shape and colors of the head are extracted from the ToF image and the color image, respectively. Then, a head model is defined with the polygon and texture information.
3. Head tracking
   Head tracking is realized by fitting the video image into the head model. A particle filter is adopted to track the three-dimensional position of the head and its three-dimensional orientation.
4. Identification of eye-gaze object
   From the six-dimensional parameters, an eye-gaze vector is computed in the three-dimensional space. The object of the eye-gaze is determined by this vector and the position of the objects. In this study, the eye-gaze object is limited to the poster and other participants.

The entire process mentioned above can be run in real time by using a GPU for tracking each person.

#### 9.2.3.2 Detection of Nodding

Nodding can be detected as a movement of the head, whose position is estimated in the above process. However, discrimination against noisy or unconscious movements is still difficult. Therefore, nodding is not used in most of this study.

#### 9.2.3.3 Speech Separation and Speaker Diarization

Speech separation and enhancement are realized with the blind spatial subtraction array (BSSA), which consists of the delay-and-sum (DS) beamformer and a noise estimator based on independent component analysis (ICA) [8]. Here, the position information of the participants estimated by the image processing is used for beamforming and initialization of the ICA filter estimation. This is one of the advantages of multi-modal signal processing. While the participants move around, the filter estimation is updated online.

When the 19-channel microphone array is used, speech separation and enhancement can be performed with a high SNR, but not in real time. Using the Kinect sensor realizes real-time processing, but degrades the quality of speech.

By this process, the audio input is separated to the presenter and the audience. Although discrimination among the audience is not done, DoA (Direction of Arrival) estimation can be used for identifying the speaker among the audience. In a baseline system, simple voice activity detection (VAD) is conducted on each of the two channels by using power and spectrum information in order to make speaker diarization. We can use highly-enhanced but distorted speech for VAD, but still keeps moderately-enhanced and intelligible speech for re-playing.

In Sect. 9.4, a more elaborate speaker diarization method is addressed by combining multi-channel audio input and eye-gaze information of the participants.

## 9.3 Prediction of Turn-Taking from Multi-modal Behaviors

Turn-taking in conversations is a natural behavior in human activities. Studies on turn-taking have been conventionally focused on dyadic conversations between two persons. While there are a number of studies conducting analysis on the turn-taking patterns [9–12], some studies investigated a prediction mechanism for a dialogue system to take or yield turns based on machine learning [13–16]. Some studies even attempt to evaluate the synchrony of dialogue [17, 18].

Recently, conversational analysis and modeling have been extended to multi-party interactions such as meetings and free conversations by more than two persons. Turn-taking in multi-party interactions is more complicated than that in the dyadic dialogue case, in which a long pause suggests yielding turns to the (only one) partner. Predicting whom the turn is yielded to or who will take the turn is significant for an intelligent conversational agent handling multiple partners [19, 20] as well as an automated system to beamform microphones or zoom in cameras on the speakers.

Studies on computational modeling on turn-taking in multi-party interactions are very limited so far. Laskowski et al. [21] presented a stochastic turn-taking model based on N-gram for the ICSI meeting corpus. Jokinen et al. [22] investigated the use of eye-gaze information for predicting turn-holding or giving in three-party conversations.

This section deals with turn-taking behaviors in poster sessions. Conversations in poster sessions are different from those in meetings and free conversations addressed in the previous works, in that presenters hold most of turns and thus the amount of utterances is very unbalanced. However, the segments of audiences' questions and comments are more informative and should not be missed, and thus prediction of such events is important in online applications such as automated recording control and a conversational agent. Therefore, the goal of this work is to predict turn-taking by the audience in poster conversations, and, if that happens, which person in the audience will take the turn to speak.

We approach this problem by combining multi-modal information sources. While most of the aforementioned previous studies focused on prosodic features of the current speakers, it is widely-known that eye-gaze information plays a significant role in turn-taking [23], and the works by Jokinen [22] and by Bohus [19] exploited that information in their modeling. The existence of posters, however, requires different modeling in poster conversations as the eye-gaze of the participants are focused on the poster in most of the time. This is true to other kinds of interactions using some materials such as maps and computers. Several kinds of parameterization of eye-gaze patterns including the poster object are investigated for effective features related with turn-taking. Moreover, backchannel information such as nodding and verbal reactions by the audience is also incorporated

In this study, four poster sessions are used. In majority of utterances (IPUs) of the presenter ("A"), the turn was held by himself/herself. The ratio of turn-taking by the audience (either "B" or "C") is only 11.9 %. In this work, therefore, prediction of turn-taking is formulated as a detection problem rather than a classification problem. The evaluation measure should be recall and precision of turn-taking by the audience, not the classification accuracy of turn-holding and yielding by the presenter. This is consistent with the goal of the study.

## *9.3.1 Analysis on Eye-Gaze and Backchannel Features in Turn-Taking*

First, statistics of eye-gaze and backchannel events are investigated on their relationship with turn-taking by the audience.

### 9.3.1.1 Distribution of Eye-Gaze

The object of the eye-gaze of all participants is identified at the end of the presenter's utterances. The target object can be either the poster or other participants. The

**Fig. 9.5** Statistics of eye-gaze and its relationship with turn-taking (ratio)

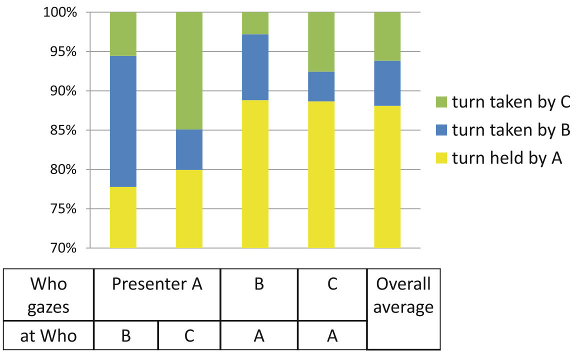

**Table 9.1** Duration of eye-gaze and its relationship with turn-taking (s)

| | Turn held by presenter | Turn taken by audience | |
|---|---|---|---|
| | A | B | C |
| A gazed at B | 0.220 | **0.589** | 0.299 |
| A gazed at C | 0.387 | 0.391 | **0.791** |
| B gazed at A | 0.161 | 0.205 | 0.078 |
| C gazed at A | 0.308 | 0.215 | 0.355 |

statistics are shown in Fig. 9.5 in relation with the turn-taking events. It is observed that the presenter is more likely to gaze at the person in the audience right before yielding the turn to him/her. We can also see that the person who takes the turn is more likely to gaze at the presenter, but the ratio of the turn-yielding by the presenter is not higher than the average over the entire data set.

The duration of the eye-gaze is also measured. It is measured within the segment of 2.5 s before the end of the presenter's utterances because the majority of the IPUs are less than 2.5 s. It is listed in Table 9.1 in relation with the turn-taking events. We can see the presenter gazed at the person right before yielding the turn to him/her significantly longer than other cases. However, there is no significant difference in the duration of the eye-gaze by the audience according to the turn-taking events.

#### 9.3.1.2 Joint Eye-Gaze Events

Next, joint eye-gaze events by the presenter and the audience are defined as shown in Table 9.2. In this table, notation of "audience" is used, but actually these events are defined for each person in the audience. Thus, "Ii" means the mutual gaze by the presenter and a particular person in the audience, and "Pp" means the joint attention to the poster object.

**Table 9.2** Definition of joint eye-gaze events by presenter and audience

| Who | Presenter | | |
|---|---|---|---|
| | Gazes at | Audience **(I)** | Poster **(P)** |
| Audience | Presenter **(i)** | **Ii** | **Pi** |
| | Poster **(p)** | **Ip** | **Pp** |

**Table 9.3** Statistics of joint eye-gaze events by presenter and audience in relation with turn-taking (ratio of occurrence frequency)

| | #Turn held by presenter A (%) | #Turn taken by audience | | Total (%) |
|---|---|---|---|---|
| | | (Self) (%) | (Other) (%) | |
| Ii | 3.1 | 0.4 | 0.1 | 3.6 |
| Ip | 7.9 | **1.8** | 0.6 | 10.3 |
| Pi | 4.7 | 0.3 | 0.2 | 5.2 |
| Pp | 73.7 | 3.6 | 3.6 | 80.9 |

Statistics of these events at the end of the presenter's utterances are summarized in Table 9.3. Here, the counts of the events are summed over the two persons in the audience. They are classified according to the turn-taking events, and turn-taking by the audience is classified into two cases: the person involved in the eye-gaze event actually took the turn (self), and the other person took the turn (other). It is confirmed that the joint gaze at the poster is most dominant (around 80 %) in the poster conversations. The mutual gaze ("Ii") is expected to be related with turn-taking, but its frequency is not so high. The frequency of "Pi" is not high, either. The most potentially useful event is "Ip", in which the presenter gazes at the person in the audience before giving the turn. This is consistent with the observation in the previous subsection.

#### 9.3.1.3 Dynamics of Eye-Gaze

In the analysis of the previous subsections, gazing information by the audience is not so clearly related with turn-taking. The audience might have sent a signal to the presenter by gazing that he would like to take a turn, but turn-taking actually happens when the presenter looks back to him/her. To confirm this, the dynamic patterns of the eye-gaze events are investigated by a window of 2.5 s over 10 s before the end of the presenter's utterances. As a result, we observe a tendency that the frequency and duration of "Ii" and "Ip" are increasing toward the end of the utterances, while "Pi" appeared relatively longer in the segment of 5 s before the end of the utterances. This indicates that "Pi" is followed by "Ii" or "Ip". This suggests that bigram information of the eye-gaze events may be useful when we have a larger amount of data.

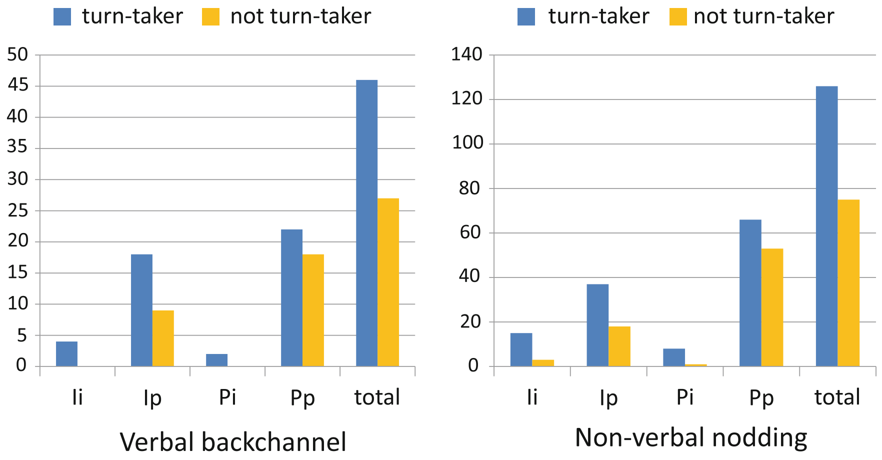

**Fig. 9.6** Statistics of backchannels and their relationship with turn-taking (occurrence frequency)

#### 9.3.1.4 Backchannels

Verbal backchannels, typically "*hai*" in Japanese and "yeah" or "okay" in English, indicate that the listener is understanding what is being said. Nodding is regarded as a non-verbal backchannel, and it is more frequently observed in poster conversations than in simple spoken dialogue.

The occurrence frequencies of these events are counted within the segment of 2.5 s before the end of the presenter's utterances. They are shown in Fig. 9.6 according to the joint eye-gaze events. It is observed that the person in the audience who takes the turn (=turn-taker) made more backchannels both in verbal and non-verbal manners, and the tendency is more apparent in the particular eye-gaze events of "Ii" and "Ip" which are closely related with the turn-taking events.

### *9.3.2 Prediction of Turn-Taking by Audience*

Based on the analysis in the previous subsection, features for predicting turn-taking by the audience are parameterized. The prediction task is divided into two sub-tasks: detection of speaker change and identification of the next speaker. In the first subtask, we predict whether the turn is yielded from the presenter to (someone in) the audience, and if that happens, then we predict who in the audience takes the turn in the second sub-task. Note that these predictions are done at every end-point of the presenter's utterance (IPU) using the information prior to the speaker change or the utterance by the new speaker.

Prediction experiments were conducted based on machine learning using the data set in a cross-validation manner; one session is tested using the classifier trained with the other sessions, and this process is repeated by changing the training and testing set.

#### 9.3.2.1 Prediction of Speaker Change

For the first sub-task, prosodic features are adopted as a baseline based on the previous works (e.g. [16, 22]). Specifically, F0 (mean, max, min, and range) and power (mean and max) of the presenter's utterance is computed prior to the prediction point. Each feature is normalized by the speaker by taking the z-score; it is subtracted by the mean and then divided by the variance for the corresponding speaker.

Backchannel features are defined by taking occurrence counts prior to the prediction point for each type (verbal backchannel and non-verbal nodding).

Eye-gaze features are defined as below:

1. Eye-gaze object
   For the presenter, (P) poster or (I) audience;
   For (anybody in) the audience, (p) poster, (i) presenter, or (o) other person in the audience.
2. Joint eye-gaze event: "Ii", "Ip", "Pi", "Pp"
   These can happen simultaneously for multiple persons in the audience, but only one is chosen by the priority order listed above.
3. Duration of the above 1. ((I) and (i))
   A maximum is taken over persons in the audience.
4. Duration of the above 2. (except "Pp")

Note that these parameters can be extended to any number of the persons in the audience, although only two persons were present in this data set.

Support vector machines (SVM) and logistic regression (MaxEnt) model are used for machine learning, but they show comparable performance. The result with SVM is listed in Table 9.4. Here, recall, precision and F-measure are computed for speaker change, or turn-taking by the audience. This case accounts for only 11.9 % and its prediction is a very challenging task, while we can easily get an accuracy of over 90 % for prediction of turn-holding by the presenter. We are particularly concerned on the recall of speaker change, considering the nature of the task and application scenarios.

Among the individual features, as shown in Table 9.4, the prosodic features obtain the best recall while the eye-gaze features achieve the best precision and F-measure. In the table, combination of all four kinds of the eye-gaze parameterization listed above is adopted, however, using one of them is sufficient and there is not a significant difference in performance among them. Combination of the prosodic features and

**Table 9.4** Prediction result of speaker change

| Feature | Recall | Precision | F-measure |
|---|---|---|---|
| Prosody | 0.667 | 0.178 | 0.280 |
| Backchannel (BC) | 0.459 | 0.113 | 0.179 |
| Eye-gaze (gaze) | 0.461 | 0.216 | 0.290 |
| Prosody+BC | 0.668 | 0.165 | 0.263 |
| Prosody+gaze | **0.706** | 0.209 | 0.319 |
| Prosody+BC+gaze | 0.678 | 0.189 | 0.294 |

**Table 9.5** Prediction result of the next speaker

| | Feature | Accuracy (%) |
|---|---|---|
| 1. | Eye-gaze object | 53.8 |
| 2. | Joint eye-gaze event | 53.8 |
| | 1.+2. | 55.8 |
| 3. | 1.+2. + duration | 66.4 |
| BC | Backchannel | 52.6 |
| | Combination of above all (3.+BC) | **69.7** |

eye-gaze features is effective in improving both recall and precision. On the other hand, the backchannel features get the lowest performance, and its combination with the other features is not effective, resulting in degradation of the performance.

#### 9.3.2.2 Prediction of Next Speaker

Predicting the next speaker in a multi-party conversation (before he/she actually speaks) is also a challenging task, and has not been addressed in the previous work. For this sub-task, the prosodic features of the current speaker are not usable because it does not have information suggesting who the turn will be yielded to. Therefore, the backchannel features and eye-gaze features described in the previous subsection are adopted, but they are computed for individual persons in the audience, instead of taking the maximum or selecting among them.

In this experiment, SVM performs slightly better than logistic regression model, thus the prediction accuracy obtained with SVM is listed in Table 9.5. As there are only two persons in the audience, random selection would give an accuracy of 50 %.

The simple eye-gaze features focused on the prediction point (1 and 2) obtains an accuracy slightly better than the chance rate, but incorporating duration information (3) significantly improves the accuracy. In this experiment, the backchannel features have some effect; the person who made more backchannels is more likely to take the turn. By combining all features, the accuracy reaches almost 70 %.

## 9.4 Speaker Diarization with Backchannel Detection Using Eye-Gaze Information

In the previous section, it is shown that eye-gaze information is useful for predicting turn-taking. Based on this finding, we investigate a new scheme of speaker diarization. Speaker diarization is a process to identify "who spoke when" in multi-party conversations. A number of diarization methods [24, 25] have been investigated based on acoustic information. In real multi-party conversations, the diarization performance is degraded by adversary acoustic conditions such as background noise and distant talking. To solve the problem, some studies tried to incorporate multi-modal information such as motion and gesture [12, 25].

Although it is known that eye-gaze information can be used to predict participants' utterances, it has not been integrated in speaker diarization tasks. This section addresses a multi-modal diarization method which integrates eye-gaze information with acoustic information. The proposed method extracts acoustic and eye-gaze features, which are integrated in a stochastic manner to detect utterances.

Furthermore, the diarization results are enhanced by detecting audience's backchannels. Backchannels are frequently observed in poster conversations and involve different eye-gaze behaviors since they indicate that the listener does not take a turn. Detection of backchannels is also realized by using the same multi-modal scheme but training a different model. By eliminating the detected backchannels and noise from the diarization result, we can easily access to meaningful utterances such as questions and comments, while backchannels show interaction level of the conversation.

In this study, eight poster sessions are used. Since utterances by the audience are not frequent, it is difficult to detect these utterances accurately. Moreover, the audience's backchannels account for about 40 % of their utterance duration.

### *9.4.1 Multi-modal Speaker Diarization*

#### 9.4.1.1 MUSIC Method Using Microphone Array

Conventional speaker diarization methods have used Mel-Frequency Cepstral Coefficients (MFCCs) and Directions Of Arrival (DOA) of sound sources [24]. An acoustic baseline method in this study is based on sound source localization using DOAs derived from the microphone array.

To estimate a DOA, we adopt the MUltiple SIgnal Classification (MUSIC) method [26], which can detect multiple DOAs simultaneously. The MUSIC spectrum $M_t(\theta)$ is calculated based on the orthogonal property between an input acoustic signal and a noise subspace. Note that $\theta$ is an angle between the microphone array and the target of estimation, and $t$ represents a time frame. The MUSIC spectrum represents DOA likelihoods, and the large spectrum suggests that the participant makes an utterance from that angle. To calculate the spectrum, it is needed to determine the number of sound sources. In this study, the number of sound sources is predicted with SVM using the eigenvalue distribution of a spatial correlation matrix [27].

The proposed method incorporates eye-gaze information to speaker diarization. The method first extracts acoustic and eye-gaze features to compute a probability of speech activity respectively, then it combines the two probabilities for the frame-wise decision. The process is conducted independently on every time frame $t$ and for each participant $i$.

The acoustic features are calculated based on the MUSIC spectrum. We can use the $i$th participant's head location $\theta_{i,t}$ tracked by the Kinect sensors. The possible location of the participant is constrained within a certain range ($\pm\theta_B$) from the detected location $\theta_{i,t}$. The acoustic features of the $i$th participant in the time frame $t$ consist of the MUSIC spectrum in the range:

$$\mathbf{a}_{i,t} = \left[M_t\left(\theta_{i,t} - \theta_B\right), \cdots, M_t\left(\theta_{i,t}\right), \cdots, M_t\left(\theta_{i,t} + \theta_B\right)\right]^T . \tag{9.1}$$

#### 9.4.1.2 Eye-Gaze Features

The eye-gaze features for the $i$th participant $\mathbf{g}_{i,t}$ are same as those used in Sect. 9.3.2.1, except that unigram and bigram of the eye-gaze objects and the joint eye-gaze events are added.

#### 9.4.1.3 Integration of Acoustic and Eye-Gaze Information

The acoustic features $\mathbf{a}_{i,t}$ are integrated with the eye-gaze features $\mathbf{g}_{i,t}$ to detect the $i$th participant's speech activity $v_{i,t}$ in the time frame $t$. Note that the speech activity $v_{i,t}$ is binary: speaking ($v_{i,t} = 1$) or not-speaking ($v_{i,t} = 0$). Here, a linear interpolation is adopted to combine probabilities independently computed by the two feature sets [25]:

$$f_{i,t}(\mathbf{a}_{i,t}, \mathbf{g}_{i,t}) = \alpha\ p(v_{i,t} = 1|\mathbf{a}_{i,t}) +\ (1-\alpha)\ p(v_{i,t} = 1|\mathbf{g}_{i,t}) . \tag{9.2}$$

Here $\alpha \in [0, 1]$ is a weight coefficient. Each probability is computed by a logistic regression model. It is also possible to combine the two feature sets in the feature domain and directly compute a posterior probability $p(v_{i,t}|\mathbf{a}_{i,t}, \mathbf{g}_{i,t})$. Compared with this joint model, the linear interpolation model has a merit that training data does not have to be aligned between the acoustic and eye-gaze features because of independency of the two discriminative models. Furthermore, the weight coefficient $\alpha$ can be appropriately determined based on the acoustic environments such as Signal-to-Noise Ratio (SNR). Here, it is estimated using an entropy $h$ of the acoustic posterior probability $p(v_{i,t}|\mathbf{a}_{i,t})$ [28] as

$$\alpha = \alpha_c \cdot \frac{1-h}{1-h_c}\ , \tag{9.3}$$

where $h_c$ and $\alpha_c$ are an entropy and an ideal weight coefficient in a clean acoustic environment, respectively. When the estimated weight coefficient is larger than one or less than zero, the coefficient is set to one or zero, respectively. For online processing, the coefficient is updated periodically.

#### 9.4.1.4 Speaker Diarization Experiment

Logistic regression models were trained separately for the presenter and the audience by cross-validation of the eight sessions. In order to evaluate performance under ambient noise, audio data was prepared by superimposing a diffusive noise recorded in a crowded place. SNRs were set to 20, 15, 10, 5 and 0 dB. In real poster conversations carried out in academic conventions, the SNRs are expected to be around 0 to 5 dB.

The multi-modal method is compared with other methods listed below:

1. *baseline MUSIC* [29]
   This method conducts peak tracking of the MUSIC spectrum and GMM-based clustering in the angle domain. Each cluster corresponds to each participant. This method does not use any cue from visual information.
2. *baseline + location constraint* [30]
   This method also performs peak tracking of the MUSIC spectrum, and compares the detected peak with the estimated head location within the $\pm\theta_B$ range. If this constraint is not met, the hypothesis is discarded.
3. *acoustic-only model*
   This method fixes the weight coefficient $\alpha$ to 1 in Eq. (9.2), and uses only the acoustic information.

For an evaluation measure, Diarization Error Rate (DER) [31] is used in this experiment. DER consists of False Acceptance (FA), False Rejection (FR), and Speaker Error (SE) as below:

$$DER = \frac{\#FA + \#FR + \#SE}{\#S}\ , \tag{9.4}$$

where $\#S$ is the number of speech frames in the reference data.

Table 9.6 lists DERs for each SNR. The two baseline methods (*baseline MUSIC* and *baseline + location constraint*) showed lower accuracy because they are rule-based and not robust against dynamic changes of the MUSIC spectrum and participants' locations. Compared with the acoustic-only model, the proposed multi-modal model achieves higher performance under noisy environments (SNR = 5, 0 dB). Thus, we can see the effect of the eye-gaze information under noisy environments expected in real poster sessions.

**Table 9.6** Evaluation of speaker diarization (DER [%])

| Method | | SNR (dB) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $\infty$ | 20 | 15 | 10 | 5 | 0 | Average |
| *Baseline MUSIC* | [29] | 16.94 | 23.14 | 31.66 | 47.92 | 67.03 | 88.80 | 45.92 |
| *Baseline + location constraint* | [30] | 8.34 | 14.45 | 22.31 | 36.09 | 55.80 | 78.05 | 35.84 |
| *Acoustic-only model* | Eq. (9.2) w/o $\mathbf{g}_{i,t}$ | **6.16** | **7.28** | **9.36** | 14.20 | 22.94 | 35.89 | 15.97 |
| *Multi-modal model* | Eq. (9.2) | 6.27 | 7.81 | 9.96 | **13.69** | **18.18** | **21.61** | **12.92** |

The weight coefficient $\alpha$ in Eq. (9.2) was also manually tuned where the stepping size was 0.1. In the clean environment (SNR = $\infty$ dB), the optimal weight was 1.0. On the other hand, in the noisy environments (SNR = 5 and 0 dB), the optimal weights were 0.6 or 0.5. These results suggest that the weight of eye-gaze features is appropriately increased in noisy environments. The average DER by the manual tuning is 11.78 %, which is slightly better than the result (12.92 %) by the automatic weight estimation (Eq. 9.3). Therefore, the automatic weight estimation works reasonably according to the acoustic environment.

### 9.4.2 *Detection of Backchannels*

The diarization result includes backchannels and also falsely accepted noise especially for audience's utterances. A post-processing model is introduced to detect and eliminate them and highlight questions and comments by the audience, which are important for efficient review of poster conversations. There have been few works on detection of backchannels while many studies have been conducted to predict appropriate timing of backchannels [32–35].

Backchannels suggest that the current speaker can hold the turn, and the listener does not take a turn. In that sense, the eye-gaze behaviors are different from those of turn-taking. Thus, a different model is trained using the eye-gaze behaviors to predict backchannels. Here, the multi-modal scheme formalized in the previous subsection is modified. The eye-gaze features and the multi-modal integration model are same, but here the acoustic features are re-designed. Multi-channel acoustic signals are enhanced for each participant by delay-and-sum beamforming. The enhanced signal is used to calculate the acoustic features as follows:

1. the number of time frames of the utterance segment calculated from the diarization result
2. MFCC parameters (12-MFCCs and 12-$\Delta$MFCCs)
3. Power (and $\Delta$Power)
4. Regression coefficients of fundamental frequency (F0) and power at the end of the preceding utterance [34]

Logistic regression models are trained to predict three events: backchannels, utterances other than backchannels, and noise. For each utterance segment as a result of speaker diarization, cumulative likelihoods are calculated by the three models, and they are normalized so that the sum of the three is one. The eliminated utterance segments are determined by the thresholding operation with a sum of the posterior probabilities on backchannels and noise.

The diarization result is post-processed by another model for elimination of backchannels and noise. The reference labels in this experiment regard backchannels as non-speech events.

The following methods are compared. They were applied after the multi-modal speaker diarization (last row of Table 9.6).

1. *thresholding with utterance duration*
   A threshold in this method is the duration of each utterance section since the duration of backchannels is usually shorter than others. This corresponds to using only the first feature listed above.
2. *acoustic-only model*
   This method uses the acoustic features listed above.
3. *multi-modal model*
   This method also uses the eye-gaze features in addition to the acoustic features.

Here, we focus on substantial utterances by the audience for efficient access to the recordings. Since there are rarely overlapping utterances other than backchannels, we measured Equal Error Rate (EER) where False Acceptance Rate (FAR) equals to False Rejection Rate (FRR). FAR and FRR are defined as:

$$FAR = \frac{\#FA}{\#NS}, \quad FRR = \frac{\#FR}{\#S}, \tag{9.5}$$

where $\#NS$ is the number of non-speech frames in the reference. EER is calculated by varying the threshold in speaker diarization.

Table 9.7 lists EERs for each SNR. Compared to the case without post-processing (*no post-processing*), the proposed multi-modal model significantly reduces EERs. This shows the effectiveness of elimination of backchannels and noise after speaker diarization. The simple thresholding method (*thresholding with utterance duration*) reduces EERs in noisy conditions, but degrades in clean conditions. It is difficult to detect backchannels only with the utterance duration. The effect of the eye-gaze features is also confirmed under noisy environments (SNR = 5.0 dB).

**Table 9.7** Evaluation of audience's speech detection (EER [%])

| Method | | SNR (dB) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $\infty$ | 20 | 15 | 10 | 5 | 0 | Average |
| *No post-processing* | | 13.37 | 15.80 | 17.86 | 20.86 | 25.77 | 31.80 | 20.91 |
| *Thresholding with utterance duration* | | 15.95 | 17.60 | 18.64 | 20.38 | 24.74 | 30.81 | 21.35 |
| *Acoustic-only model* | Eq. (9.2) w/o $\mathbf{g}_{i,t}$ | **12.14** | **13.98** | 15.47 | **18.19** | 23.34 | 30.20 | 18.89 |
| *Multi-modal model* | Eq. (9.2) | 12.23 | 14.11 | **15.42** | 18.29 | **23.07** | **29.72** | **18.80** |

## 9.5 Detection of Hot Spots via Prominent Reactive Tokens of Audience

This section addresses high-level indexing of poster conversations based on the interactive characteristics. As opposed to the conventional content-based approach which focuses on the presenter's speech, we focus on the audience's reaction, specifically the audience's reactive tokens and laughter. By reactive tokens (*Aizuchi* in Japanese), we mean the listener's verbal short response, which expresses his/her state of the mind during the conversation. We particularly focus on prominent non-lexical reactive tokens, such as "*hu:n*", "*he:*" in Japanese and "wow", "gosh" in English, which are not used for simple acknowledgment and presumably related with the state of the mind of the listener. These can be articulated with a variety of prosodic patterns; they can be prolonged to an arbitrary length.

It is assumed that the audience signals their interest level with this kinds of non-lexical reactive tokens, and that detection of the audience's interest level is useful for indexing the speech archives, because people would be interested in listening to the points other people were interested in. It is also presumed that people would be interested in the funny spots where laughter was made. In this work, those spots which induced (or elicited) laughter and non-lexical reactive tokens are defined as hot spots, and their automatic detection is investigated.

In this study, eight poster sessions are used.

### *9.5.1 Detection of Laughter and Reactive Tokens*

Detection of laughter has been addressed by several studies [36–38]. Typically, a dedicated classifier such as GMM and SVM is prepared for discriminating laughter against speech. On the other hand, studies on detecting reactive tokens is limited. Ward [39] investigated prosodic patterns of reactive tokens, but did not conduct automatic detection. Other works [40, 41] focused on distinction of affirmative answers "yes" and tokens used in backchannels. In Japanese, there are a variety of syllabic patterns in reactive tokens, including both lexical and non-lexical tokens.

A framework for acoustic event detection in audio recordings of conversations is designed based on a combination of BIC-based segmentation and GMM-based classification [42]. For each segment, classification based on GMM is applied. GMMs are prepared for five classes of male speech, female speech, noise, laughter and reactive tokens. Laughter is detected with this GMM-based classification.

Reactive tokens are more difficult to detect, because they are much similar to normal speech in terms of acoustic characteristics. Thus, we incorporate two additional processes to verify the candidates of reactive tokens hypothesized by GMM-based classification. One is the filled pause detector which considers monotonousness of spectral and pitch patterns [43]. The other is a speech recognition system, which is used to filter out filled pauses included in its lexicon. In summary, reactive tokens are detected only when supported by the following three classifiers.

- dedicated GMM
- filled pause detector (to reject normal speech)
- speech recognizer (to reject fillers)

Detection accuracy of laughter and reactive tokens is shown in Table 9.8 with evaluation measures of recall, precision and F-measure. Here, F-measure is defined with a double weight on precision, because there are a number of indistinct laughter and reactive tokens, which are hard to recall and not useful for indexing.

As shown in Table 9.8, overall recall is not high, but we can detect most of the distinct events such as loud laughter and long reactive tokens. These distinct events are more related with the hot spots than subtle events. The frame-wise classification accuracy among five GMM classes is 82.3 %.

### 9.5.2 *Subjective Evaluation of Detected Hot Spots*

Based on the detected laughter and reactive tokens, hot spots are defined to correspond to these two kinds of events. Specifically, hot spots are labeled for utterances which induce (or elicit) the events. The segments are defined by utterance units, i.e. made of a couple of utterances, with a maximum duration determined by a threshold.

**Table 9.8** Detection accuracy of laughter and reactive tokens

| | Recall | Precision | F-measure |
|---|---|---|---|
| Laughter | 0.419 | 0.750 | 0.648 |
| Reactive token | 0.439 | 0.707 | 0.630 |

**Table 9.9** Ratio of appropriate hot spots among detected spots ("precision")

| | Precision (oracle) |
|---|---|
| Spots accompanying laughter | 74.7 % (89.2 %) |
| Spots accompanying reactive token | 86.5 % (95.2 %) |

Subjective evaluations were conducted on the hot spots indexed in this manner. Four subjects, who had not attended the presentation nor listened to the recorded audio content, were asked to listen to each of the segmented hot spots in the original time sequence, and to make evaluations on the questionnaire, as below.

Q1: Do you understand the reason why the reactive token/laughter occurred?
Q2: Do you find this segment interesting/funny?
Q3: Do you think this segment is necessary or useful for listening to the content?

The result of Question 1 (percentage of "yes"), summarized in Table 9.9, suggests the ratio of appropriate hot spots or "precision" among the detected hot spots, because the third person verified the spots were naturally inducing laughter or reactive tokens. The figures labeled "(oracle)" in Table 9.9 show the result when limited to the segments where laughter or reactive tokens were correctly detected. It is confirmed that a large majority of the detected spots are appropriate. There are more "false" detections for the segments accompanying laughter; laughter is socially made to relax the participants in the poster conversations.

The answers to Questions 2 and 3 are more subjective, but suggest the usefulness of the hot spots. Only a half of the spots associated with laughter are funny for the subjects (Q2), and they found 35 % of the spots not funny. The result suggests that feeling funny largely depends on the person. And we should note that there are not many funny parts in the poster sessions by nature.

On the other hand, more than 90 % of the spots associated with reactive tokens are interesting (Q2), and useful or necessary (Q3) for the subjects. The result supports the effectiveness of the hot spots extracted based on the reaction of the audience.

### *9.5.3 Prosodic Analysis of Reactive Tokens*

In the system described above, all non-lexical reactive tokens are detected without considering their syllabic and prosodic patterns. In this subsection, syllabic and prosodic patterns of reactive tokens related with the interest level are investigated Generally, prosodic features play an important role in conveying para-linguistic and non-verbal information. In previous works [40, 41], it was reported that prosodic features are useful in identifying reactive tokens. Ward [39] made an analysis of pragmatic functions conveyed by the prosodic features in English non-lexical tokens.

An experiment was designed to identify the syllabic and prosodic patterns closely related with the interest level for detection of hot spots. For this investigation, three

**Table 9.10** Significant combinations of syllabic and prosodic patterns of reactive tokens

| | | Interest | Surprise |
|---|---|---|---|
| *hu:N* | Duration | * | * |
| | F0 max | | |
| | F0 range | | |
| | Power | | |
| *he:* | Duration | * | * |
| | F0 max | * | * |
| | F0 range | | * |
| | Power | * | * |
| *a:* | Duration | | |
| | F0 max | * | |
| | F0 range | | |
| | Power | * | |

syllabic patterns of "*hu:N*", "*he:*" and "*a:*" were selected. They are presumably related with the interest level and also most frequently observed in the corpus, except lexical tokens.

Duration, F0 (maximum and range) and power (maximum) are computed for each reactive token, and they are normalized for every person; for each feature, we compute the mean, and this mean is subtracted from the feature values.

For each syllabic kind of reactive token and for each prosodic feature, top-ten and bottom-ten samples, i.e. samples that have largest/smallest values of the prosodic feature, were selected. For each of them, an audio segment was extracted to cover the reactive token and its preceding utterances. This process is similar to the hot spot detection described in the previous subsection, but was done manually according to the criteria.

Then, five subjects listened to the audio segments and evaluated the audience's state of the mind. Twelve items were evaluated in a scale of four ("strongly feel" to "do not feel"). Among them two items are related to the interest level and other two items are related to the surprise level.[1] Table 9.10 lists the results (marked by "*") that have a statistically significant ($p < 0.05$) difference between top-ten and bottom-ten samples. It is observed that prolonged "*hu:N*" means interest and surprise while "*a:*" with higher pitch or larger power means interest. On the other hand, "*he:*" can be emphasized in all prosodic features to express interest and surprise.

Using this prosodic information will enhance the precision of the hot spot detection. The tokens with larger power and/or a longer duration is apparently easier to detect than indistinct tokens, and they are more related with the hot spot. This simple principle is consistent with the proposed scheme.

[1] We used different Japanese wording for interest and for surprise to enhance the reliability of the evaluation; we adopt the result if the two matches.

## 9.6 Prediction of Interest and Comprehension Level via Audience's Questions from Multi-modal Behaviors

Feedback behaviors of an audience are important cues in analyzing presentation-style conversations. We can guess whether the audience is attracted to the presentation by observing their feedback behaviors. This characteristic is more prominent when the audience is smaller; the audience can make not only non-verbal feedbacks such as nodding, but also verbal backchannels. Eye-gaze behaviors also becomes more observable. In poster conversations, moreover, the audience can ask questions even during the presentation. By observing their reactions, particularly the quantity and quality of their questions and comments, we can guess whether the presentation is understood or liked by the audience.

In the previous section, it is shown that non-lexical reactive tokens are a good indicator of the audience's interest level. The relationship between the audience's turn-taking and feedback behaviors including backchannels and eye-gaze patterns is also confirmed.

This section addresses estimation of the interest and comprehension level of the audience based on the multi-modal behaviors. As annotation of the interest and comprehension level is apparently difficult and largely subjective, we turn to speech acts which are observable and presumably related with these mental states. One is prominent reactive tokens signaled by the audience and the other is questions raised by them. Moreover, questions are classified into confirming questions and substantive questions. Prediction of these speech acts from the multi-modal behaviors is expected to approximate the estimation of the interest and comprehension level.

In this study, ten poster sessions are used. Each poster was designed to introduce research topics of the presenter to researchers or students in other fields. It consists of four or eight components (hereafter called "slide topics") of rather independent topics. This design is a bit different from typical posters presented in academic conferences, but makes it straightforward to assess the interest and comprehension level of the audience for each slide topic. Usually, a poster conversation proceeds with an explanation of slide topics one by one, and is followed by an overall QA and discussion phase. In the QA/discussion phase, it is difficult to annotate which topic they refer. Therefore, the conversation segments of the explanation on the slide topics are used.

In the ten sessions used in this study, there are 58 slide topics in total. Since two persons participated as an audience in each session, there are 116 slots (hereafter called "topic segments") for which the interest and comprehension level should be estimated.

## 9.6.1 *Definition of Interest and Comprehension Level*

In order to get a gold-standard annotation, it would be a natural way to ask every participant of the poster conversations on the interest and comprehension level on each slide topic after the session. However, this is not possible in a large scale and also for the previously recorded sessions. The questionnaire results may also be subjective and difficult to assess the reliability.

Therefore, we focus on observable speech acts which are closely related with the interest and comprehension level. In the previous section, we identified particular syllabic and prosodic patterns of reactive tokens ("*he:*", "*a:*", "*fu:N*" in Japanese, corresponding to "wow" in English) signal interest of the audience [44]. We refer to them as prominent reactive tokens.

We also empirically know that questions raised by the audience signal their interest; the audience ask more questions to know more and better when they are more attracted to the presentation. Furthermore, we can judge the comprehension level by examining the kind of questions; when the audience asks something already explained, they must have a difficulty in understanding it.

### 9.6.1.1 Annotation of Question Type

Questions are classified into two types: confirming questions and substantive questions. The confirming questions are asked to make sure of the understanding of the current explanation, thus they can be answered simply by "Yes" or "No". [2] The substantive questions, on the other hand, are asking about what was not explained by the presenter, thus they cannot be answered by "Yes" or "No" only; an additional explanation is needed. Substantial questions are occasionally comments even in a question form.

### 9.6.1.2 Relationship Between Question Type and Interest and Comprehension Level

In four sessions, audience subjects were asked to answer their interest and comprehension level on each slide topic after the session. These are used for analysis on the relationship between these gold-standard annotations and observed questions.

Figure 9.7 shows distributions of the interest and comprehension level for each question type. The interest level is quantized into five levels from 1 (not interested) to 5 (very interested), and the comprehension level is marked from 1 (did not understand) to 5 (fully understood). In the graph, a majority of confirming questions (86 %) indicate a low comprehension level (level 1 and 2). We also see a general tendency that occurrence of questions of either types is correlated with a higher interest level (level 4&5).

[2] This does not mean the presenter actually answered simply by "Yes" or "No".

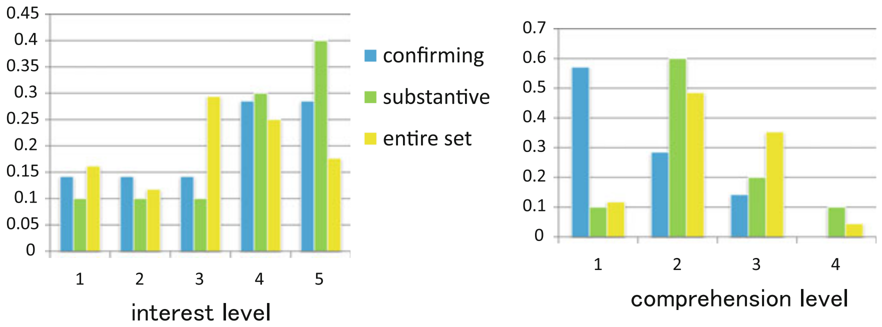

**Fig. 9.7** Distribution of interest and comprehension level according to question type

From these observations and the previous finding, the following annotation scheme is adopted.

- high interest level ← questions of any types and/or prominent reactive tokens.
- low comprehension level ← confirming questions.

Detection of these states would be particularly useful in reviewing the poster sessions or improving the presentations.

## *9.6.2 Relationship Between Multi-modal Behaviors and Questions*

Next, statistics of backchannel and eye-gaze behaviors of the audience are investigated on their relationship with questions asked by them.

### 9.6.2.1 Backchannels

It is assumed that the listener tends to make backchannels more frequently when they are attracted. In this analysis, non-lexical reactive tokens (e.g. "wow") are excluded since the prominent part of them are used for the annotation, though their occurrence frequency is much smaller (less than 20 % of all) than that of the lexical tokens (e.g. "yeah" and "okay").

Nodding is regarded as a non-verbal backchannel, and it is more frequently observed in poster conversations than in daily conversations. Our preliminary analysis showed, however, that there is not a distinct tendency in the occurrence frequency of non-verbal noddings, thus they are not used.

The occurrence frequency of the verbal backchannels normalized by the presenter's utterance (sentence unit) is counted within the topic segments. The statistics are listed according to the question type in Table 9.11. In the table, "entire" means

**Table 9.11** Relationship of audience's backchannel (count/utterance) and questions (by type)

| | Confirming | Substantive | Entire |
|---|---|---|---|
| Backchannel | 0.42 | 0.52 | 0.34 |

**Table 9.12** Relationship of audience's eye-gaze at the presenter (count/utterance and duration ratio) and questions (by type)

| | Confirming | Substantive | Entire |
|---|---|---|---|
| Gaze occurrence | **0.38** | **1.02** | 0.64 |
| Gaze duration | 0.05 | **0.15** | 0.07 |

the overall average computed for the entire topic segments of the data set. Since no questions were made in more than a half topic segments, the entire average is lower than the values in the other two columns. It is observed that the audience make more backchannels when asking questions, especially substantive questions.

### 9.6.2.2 Eye-Gaze at Presenter

The object and the duration of the eye-gaze of all participants during the topic segments are identified prior to the audiences' questions. The target object can be either the poster or other participants. In poster conversations, unlike daily conversations, participants look at the poster in most of the time. Therefore, eye-gaze at other participants has a reason and effect. The analysis in Sect. 9.3 showed that eye-gaze information is related with turn-taking events; specifically, the eye-gaze by the presenter mostly controls the turn-taking.

In this work, the eye-gaze by the audience is investigated on its relationship with the questions they ask. In particular, the eye-gaze of each person of the audience at the presenter is counted. The average occurrence count (per presenter's utterance) and the total duration (normalized per second) within the topic segments are measured. Their statistics are listed in Table 9.12. We can see a significant decrease and increase when asking confirming questions and substantive questions, respectively. It is reasoned that the audience is more focused on the poster trying to understand the content before asking confirming questions, while they want to attract the presenter's attention before asking substantive questions.

In a more detailed analysis done sentence by sentence, a gradual increase of the eye-gaze at the presenter is observed prior to substantive questions, while there is no such dynamic changes in the case of confirming questions.

The results suggest that eye-gaze information is potentially useful for identifying the question type and also estimating the interest and comprehension level.

### 9.6.3 Prediction of Interest and Comprehension Level

Based on the analysis in the previous subsection, we have implemented and evaluated classifiers to predict the interest and comprehension level of the audience in each topic segment.

First, each of audience behaviors needs to be parameterized. The features described in the previous subsection are used. An average count of backchannels per the presenter's utterance is computed. Eye-gaze at the presenter is parameterized into an occurrence count per the presenter's utterance and the duration ratio within the topic segment.

Then, regarding the machine learning method for classification, a naive Bayes classifier is adopted, as the data size is not so large to estimate extra parameters such as weights of the features. For a given feature vector $F = \{f_1, \ldots, f_d\}$, a naive Bayes classification is done by

$$p(c|F) = p(c) * \prod_i p(f_i|c)$$

where $c$ is a considered class ("high interest level or not" and "low comprehension level or not"). For computation of $p(f_i|c)$, we adopt a simple histogram quantization, in which feature values are classified into one of bins, instead of assuming a probabilistic density function. This also circumvents estimation of any model parameters. The feature bins are defined by simply splitting a histogram into 3 or 4. Then, the relative occurrence frequency in each bin is transformed into the probability form.

Experimental evaluations were done by cross-validation.

#### 9.6.3.1 Prediction of Questions and Reactive Tokens for Interest Level Estimation

First, an experiment of estimating the interest level of the audience was conducted. This problem is formulated by predicting the topic segment in which questions and/or prominent reactive tokens are made by the audience. These topic segments are regarded as "interesting" to the person who made such speech acts.

The results with different sets of features are listed in Table 9.13. F-measure is a harmonic mean of recall and precision of "interesting" segments, though recall and precision are almost same in this experiment. Accuracy is a ratio of correct output among all 116 topic segments. The chance-rate baseline when we count all segments as "interesting" is 49.1 %.

Incorporation of the backchannel and eye-gaze features significantly improves the accuracy, and the combination of both features results in the best accuracy of over 70 %. It turned out that the two kinds of parameterization of the eye-gaze feature (occurrence count and duration ratio) are redundant because dropping one of them

**Table 9.13** Prediction result of topic segments involving questions and/or reactive tokens

| | F-measure | Accuracy (%) |
|---|---|---|
| Baseline (chance rate) | 0.49 | 49.1 |
| (1) Backchannel | 0.59 | 55.2 |
| (2) Gaze occurrence | 0.63 | 61.2 |
| (3) Gaze duration | 0.65 | 57.8 |
| Combination of (1)–(3) | 0.70 | 70.7 |

**Table 9.14** Identification result of confirming or substantive questions

| | Accuracy (%) |
|---|---|
| Baseline (chance rate) | 51.3 |
| (1) Backchannel | 56.8 |
| (2) Gaze occurrence | 75.7 |
| (3) Gaze duration | 67.6 |
| Combination of (1)–(3) | 75.7 |

does not degrade the performance. However, we confirm the multi-modal synergetic effect of the backchannel and eye-gaze information.

#### 9.6.3.2 Identification of Question Type for Comprehension Level Estimation

Next, an experiment of estimating the comprehension level of the audience was conducted. This problem is formulated by identifying the confirming question given a question, which signals that the person does not understand the topic segment. Namely, these topic segments are regarded as “low comprehension (difficult to understand)” for the person who made the confirming questions.

The classification results of confirming questions versus substantive questions are listed in Table 9.14. In this task, the chance-rate baseline based on the prior statistic $p(c)$ is 51.3 %.

All features have some effects in improving the accuracy, but the eye-gaze occurrence count alone achieves the best performance and combining it with other features does not give an additional gain. This is explained by a large difference in its value among the question types as shown in Table 9.12.

As the simple occurrence frequency of backchannels is not useful for this task, the syllabic or prosodic patterns of the backchannels [45] should be investigated in the future.

## 9.7 Poster Session Browser

Based on the result and findings of this study, a poster session browser is designed and developed, as shown in Fig. 9.8. The browser visualizes activities during the poster session including speech utterances and eye-gaze at other participants for each person. It also plays the recorded audio and video based on the indices.

Along the timeline, utterance segments of each participant are marked as a result of speaker diarization and backchannel detection. We can easily access to substantial utterances from the audience such as questions and comments. Moreover, eye-gaze events are also visualized so we can estimate the interaction level of the conversation. For each person in the audience, the marked segments represent when the person gave his/her eye-gaze to the presenter.

Under the timeline, a scale-downed timeline overview is shown to allow users to outlook the entire session. By clicking a segment on the timeline overview, users can directly move to the area and see the conversation segment in the area. Poster sessions generally last very long and presenters need to explain the same content repeatedly while substantial utterances such as questions and comments by an audience is occasional but important. The above functions allow the users to efficiently access to the substantial utterances without watching the entire video.

The browser will be helpful for the presenter to review the session afterwards, since the presenter can hardly memorize the audience's questions and comments during the long session. The browser will also be useful for the colleagues or supervisor of the presenter to see how many people came to the poster and if they were interested in the presentation. It is also possible to quickly view what the audience said and how the presenter responded to them. In the future, the browser may be used in public, so viewers see the other participants' comments. But this needs to obtain a permission from the participants as well as the session organizer.

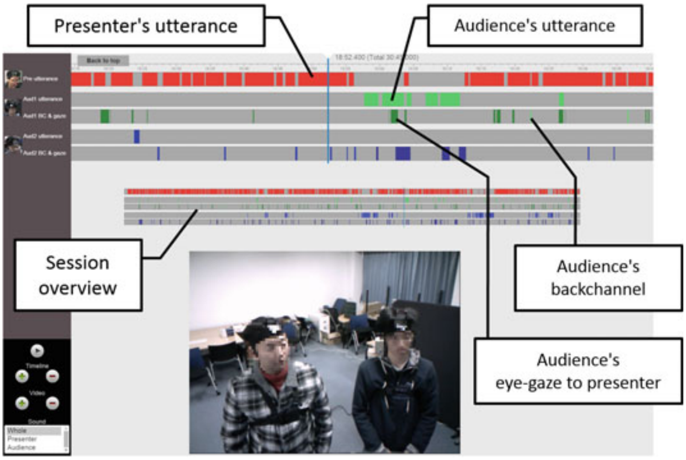

**Fig. 9.8** Poster conversation browser

**Table 9.15** Browsing time to complete quizzes; "reduction ratio" is measured against the session duration

| Subject | Time | Reduction ratio (%) |
|---|---|---|
| A | 9 m 50 s | 33.3 |
| B | 8 m 11 s | 27.7 |
| C | 8 m 13 s | 27.8 |
| D | 6 m 58 s | 23.6 |
| Average | 8 m 18 s | 28.1 |

Since the system is independent of conversational content (e.g. audio, video, utterance segment), users can easily customize this tool to other conversational forms such as meetings and discussions. Detailed information of the visualized data is described in a configuration csv file. Various types of time-series multi-modal data can be displayed on the timeline by editing the csv file. The csv file also describes the display format of the browser: colors of segments on the timeline and display positions of the visualized data. The system is designed as a Web application where the backend is implemented in Java, and the interface is implemented in HTML, CSS, and Javascript. Playing videos and audios is realized by HTML5. The browser is lightweight and OS-independent.

A simple evaluation of the browser interface was conducted by measuring the time needed for reviewing substantial exchanges in a poster session. One session was chosen from our corpus and four subjects were engaged in this experiment. They were asked to answer twelve quizzes by browsing the recorded session. The quizzes were chosen from the questions uttered by the audience and the two possible answers were prepared. The subjects were asked to select one of them which was actually given by the presenter. The questions were sorted in a time-wise random manner.

Table 9.15 shows the time the subjects expended to complete all quizzes. All subjects were able to correctly answer all quizzes in less than ten minutes, whereas the session actually lasted 29 min. On average, reviewing time is approximately 28.1 % of the duration of the session. The browser with the speaker diarization result provides an effective interface to efficiently search substantial utterances in the session.

## 9.8 Conclusions

We have conducted multi-modal conversation analysis focused on poster sessions. Poster conversations are interactive, but often long and redundant. Therefore, simple recording of the session is not so useful.

The primary goal of the study was robust signal-level sensing of participants, i.e. who came to the poster, and their verbal feedbacks, i.e. what they said. This is still challenging given distant and low-resolution sensing devices. Combination of multi-modal information sources was investigated to enhance the performance.

First, multi-modal behaviors prior to turn-taking events were investigated. For prediction of speaker change or turn-taking by the audience, both prosodic features of the presenter and eye-gaze features of all participants are useful. The most relevant among the eye-gaze information is the presenter's gazing at the speaker to whom the turn is to be yielded.

Based on this finding, a multi-modal speaker diarization method was realized by integrating eye-gaze information with acoustic information. Moreover, the diarization result was enhanced by eliminating backchannels and falsely accepted noise. The stochastic multi-modal scheme improved the performance of speaker diarization and the effect of eye-gaze information was confirmed under noisy environments.

The next step was high-level indexing of interest and comprehension level of the audience. The problem was formulated via relevant speech acts using non-verbal feedback behaviors of the audience. Two approaches were presented in this work.

One is indexing of hot spots based on the reaction of the audience, specifically, laughter and non-lexical reactive tokens. Detection of laughter is relatively easier, but the detected spots are not necessarily funny or useful, because the evaluation is largely affected by subjects. On the other hand, the spots associated with reactive tokens are consistently interesting and meaningful. Furthermore, the specific prosodic patterns closely related with the interest level were identified.

The other approach is estimation of interest and comprehension level based on the audience's feedback behaviors and speech acts such as questions and prominent reactive tokens. Specifically, estimation of the interest level was reduced to prediction of occurrence of questions and prominent reactive tokens, and estimation of comprehension level was realized by classification of the question type.

To visualize these detected events and indices, a poster session browser has been developed. The browser will be useful for assessing the effect of the processes and further improving them.

**Acknowledgments** This work was conducted by the members of the CREST project including Hiromasa Yoshimoto, Tony Tung, Yukoh Wakabayashi, Kouhei Sumi, Zhi-Qiang Chang, Takuma Iwatate, Soichiro Hayashi, Koji Inoue, Katsuya Takanashi (Kyoto University) and Yuji Onuma, Shunsuke Nakai, Ryoichi Miyazaki, Hiroshi Saruwatari (Nara Institute of Science and Technology).

## References

1. S. Renals, T. Hain, H. Bourlard, Recognition and understanding of meetings: The AMI and AMIDA projects. *Proceedings of IEEE Workshop Automatic Speech Recognition & Understanding* (2007)
2. K. Ohtsuka, Conversation scene analysis. Signal Process. Magaz. **28**(4), 127–131 (2011)
3. T. Kawahara, Multi-modal sensing and analysis of poster conversations toward smart posterboard. In *Proceedings of SIGdial Meeting Discourse and Dialogue*, pp. 1–9 (keynote speech) (2012)
4. T. Kawahara, Smart posterboard: Multi-modal sensing and analysis of poster conversations. In *Proceedings of APSIPA ASC*, page (plenary overview talk) (2013)

5. T. Kawahara, H.Setoguchi, K. Takanashi, K.Ishizuka, S. Araki, Multi-modal recording, analysis and indexing of poster sessions. *Proceedings of INTERSPEECH*, pp. 1622–1625 (2008)
6. K. Maekawa, Corpus of spontaneous Japanese: its design and evaluation. *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12 (2003)
7. H. Yoshimoto, Y. Nakamura, Cubistic representation for real-time 3D shape and pose estimation of unknown rigid object. *Proceedings ICCV, Workshop*, pp. 522–529 (2013)
8. Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, K. Shikano, Blind spatial subtraction array for speech enhancement in noisy environment. IEEE Trans. Audio, Speech Language Process. **17**(4), 650–664 (2009)
9. T. Ohsuga, M. Nishida, Y. Horiuchi, A. Ichikawa, Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue. *Proceedings INTERSPEECH*, pp. 33–36 (2005)
10. C.T. Ishi, H. Ishiguro, N. Hagita, Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts. *Proceedings of INTERSPEECH*, pp. 2006–2009 (2006)
11. N.G. Ward, Y.A. Bayyari, A case study in the identification of prosodic cues to turn-taking: back-channeling in Arabic. *Proceedings of INTERSPEECH*, pp. 2018–2021 (2006)
12. B. Xiao, V. Rozgic, A. Katsamanis, B.R. Baucom, P.G. Georgiou, S. Narayanan, Acoustic and visual cues of turn-taking dynamics in dyadic interactions. *Proceedings of INTERSPEECH*, pp. 2441–2444 (2011)
13. R. Sato, R. Higashinaka, M. Tamoto, M. Nakano, K. Aikawa, Learning decision trees to determine turn-taking by spoken dialogue systems. *Proceedings of ICSLP*, pp. 861–864 (2002)
14. D. Schlangen, From reaction to prediction: experiments with computational models of turn-taking. *Proceedings INTERSPEECH*, pp. 2010–2013 (2006)
15. A. Raux, M. Eskenazi, A finite-state turn-taking model for spoken dialog systems. *Proceedings of HLT/NAACL* (2009)
16. N.G. Ward, O. Fuentes, A. Vega, Dialog prediction for a general model of turn-taking. *Proceedings of INTERSPEECH*, pp. 2662–2665 (2010)
17. S. Benus, Are we 'in sync': turn-taking in collaborative dialogues. *Proceedings of INTERSPEECH*, pp. 2167–2170 (2009)
18. N. Campbell, S. Scherer, Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. *Proceedings of INTERSPEECH*, pp. 2546–2549 (2010)
19. D. Bohus, E. Horvitz, Models for multiparty engagement in open-world dialog. *Proceedings of SIGdial* (2009)
20. S. Fujie, Y. Matsuyama, H. Taniyama, T. Kobayashi, Conversation robot participating in and activating a group communication. *Proceedings of INTERSPEECH*, pp. 264–267 (2009)
21. K. Laskowski, J. Edlund, M. Heldner, A single-port non-parametric model of turn-taking in multi-party conversation. *Proceedings of ICASSP*, pp. 5600–5603 (2011)
22. K. Jokinen, K. Harada, M. Nishida, S. Yamamoto, Turn-alignment using eye-gaze and speech in conversational interaction. *Proceedings of InterSpeech*, pp. 2018–2021 (2011)
23. A. Kendon, Some functions of gaze direction in social interaction. Acta Psychol. **26**, 22–63 (1967)
24. S.E. Tranter, D.A. Reynolds, An overview of automatic speaker diarization systems. IEEE Trans. ASLP **14**(5), 1557–1565 (2006)
25. G. Friedland, A. Janin, D. Imseng, X. Anguera Miro, L. Gottlieb, M. Huijbregts, M.T. Knox, O. Vinyals, The ICSI RT-09 speaker diarization system. IEEE Trans. ASLP **20**(2), 371–381 (2012)
26. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
27. K. Yamamoto, F. Asano, T. Yamada, N. Kitawaki, Detection of overlapping speech in meetings using support vector machines and support vector regression. IEICE Trans. **E89-A**(8), 2158–2165 (2006)
28. H. Misra, H. Bourlard, V. Tyagi, New entropy based combination rules in hmm/ann multi-stream asr. Proc. ICASSP **2**, 741–744 (2003)

29. S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, S. Makino, A DOA based speaker diarization system for real meetings. *Prooceedings of HSCMA*, pp. 29–32 (2008)
30. Y. Wakabayashi, K. Inoue, H. Yoshimoto, T. Kawahara, Speaker diarization based on audio-visual integration for smart posterboard. *Proceedings of APSIPA ASC* (2014)
31. J.G. Fiscus, J. Ajot, M. Michel, J.S. Garofolo, *The Rich Transcription 2006 Spring Meeting Recognition Evaluation* (Springer, 2006)
32. H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, Y. Den, An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. Language & Speech **41**(3–4), 295–321 (1998)
33. N. Ward, W. Tsukahara, Prosodic features which cue back-channel responses in English and Japanese. J. Pragmatics **32**(8), 1177–1207 (2000)
34. N. Kitaoka, M. Takeuchi, R. Nishimura, S. Nakagawa, Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. J. Japn. Soc. Artific. Intell. **20**(3), 220–228 (2005)
35. D. Ozkan, L.-P. Morency, Modeling wisdom of crowds using latent mixture of discriminative experts. *Proceedings of ACL/HLT* (2011)
36. L.S.Kennedy, D.P.W. Ellis, Laughter detection in meetings. *NIST Meeting Recognition Workshop* (2004)
37. K.P. Truong, D.A. van Leeuwen, Automatic detection of laughter. *Proceedings InterSpeech*, pp. 485–488 (2005)
38. K.Laskowski, Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings. *Proceedings of IEEE-ICASSP*, pp. 4765–4768 (2009)
39. N. Ward, Pragmatic functions of prosodic features in non-lexical utterances. *Speech Prosody*, pp. 325–328 (2004)
40. F. Yang, G. Tur, E. Shriberg, Exploiting dialog act tagging and prosodic information for action item identification. *Proceedings of IEEE-ICASSP*, pp. 4941–4944 (2008)
41. A. Gravano, S. Benus, J. Hirschberg, S. Mitchell, I. Vovsha, Classification of discourse functions of affirmative words in spoken dialogue. *Proceedings of InterSpeech*, pp. 1613–1616 (2007)
42. K. Sumi, T. Kawahara, J. Ogata, M. Goto, Acoustic event detection for spotting hot spots in podcasts. *Proceedings of INTERSPEECH*, pp. 1143–1146 (2009)
43. M. Goto, K. Itou, S. Hayamizu, A real-time filled pause detection system for spontaneous speech recognition research. *Proceedings of EuroSpeech*, pp. 227–230 (1999)
44. T. Kawahara, Z.Q. Chang, K. Takanashi, Analysis on prosodic features of Japanese reactive tokens in poster conversations. *Proceedings Int'l Conference Speech Prosody* (2010)
45. S. Strombergsson, J. Edlund, D. House, Prosodic measurements and question types in the spontal corpus of Swedish dialogues. *Proceedings of InterSpeech* (2012)