SIP (2016), vol. 5, e2, page 1 of 12 © The Authors, 2016.

doi:10.1017/ATSIP.2016.2

ORIGINAL PAPER

Multi-modal sensing and analysis of poster conversations with smart posterboard

TATSUYA KAWAHARA, TAKUMA IWATATE, KOJI INOUE, SOICHIRO HAYASHI, HIROMASA YOSHIMOTO AND KATSUYA TAKANASHI

Conversations in poster sessions in academic events, referred to as poster conversations, pose interesting, and challenging topics on multi-modal signal and information processing. We have developed a smart posterboard for multi-modal recording and analysis of poster conversations. 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 incorporates face and eye-gaze tracking for effective speaker diarization. It is demonstrated 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 signal processing, Conversation analysis, Behavioral analysis, Speaker diarization

Received 27 May 2015; Revised 31 January 2016

I. INTRODUCTION

Multi-modal signal and information processing has been investigated primarily for intelligent human–machine interfaces, including smart phones, KIOSK terminals, and humanoid robots. Meanwhile, 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. In this scenario, sensing devices are installed in an ambient manner. Examples of this kind of direction include meeting capturing [1] and conversation analysis [2].

We have been conducting a project which focuses 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.

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 this study 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.

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 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. 1.

Academic Center for Computing and Media Studies, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

Corresponding author:
T. Kawahara
Email: kawahara@i.kyoto-u.ac.jp

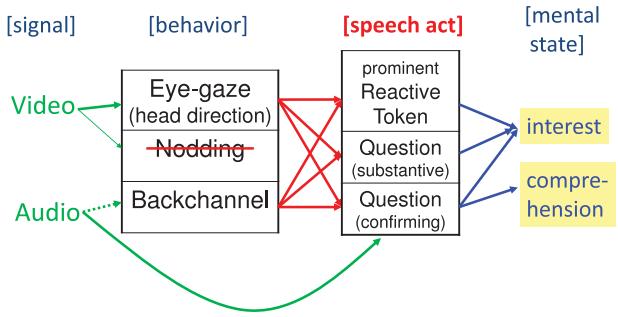

Fig. 1. Proposed scheme of multi-modal sensing and analysis.

The primary technical contribution of this paper is to figure out the effect of eye-gaze information on prediction of turn-taking events and then on speaker diarization in multi-party conversations. The secondary novelty is to investigate the relationship between eye-gaze information and the speech acts which are related with the interest and comprehension level. In this work, backchannel information is not addressed. It was addressed in our previous work [5].

In the remainder of the paper, the steps in Fig. 1 are explained after a brief description of the multi-modal corpus in Section II. Section III gives a process overview of audio-visual sensing of conversation participants and their speech using the ambient devices installed in the posterboard (green lines). In Section IV, the relationship between the eye-gaze events and turn-taking is analyzed. In Section V, a novel multi-modal speaker diarization method using eye-gaze information is presented and evaluated. In Section VI, speech acts and mental states of the audience is analyzed to define the interest and comprehension level (blue lines), and then prediction of the concerned speech acts from the audience's multi-modal behaviors (red lines) is addressed.

II. MULTI-MODAL CORPUS OF POSTER CONVERSATIONS

We have recorded a number of poster conversations for multi-modal interaction analysis [3, 6]. 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.

For the ground-truth annotation, special multi-modal sensing devices such as a motion capturing system were used, while every participant wore a wireless head-set microphone and an eye-tracking recorder. 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. The detected eye-gaze events are verified by a human annotator. We also introduced the magnetometric sensor, worn by every participant, to measure the head position and orientation accurately.

All speech data were segmented into Inter-Pausal Unit (IPUs) and sentence units with time and speaker labels, and transcribed according to the guideline of the Corpus of Spontaneous Japanese (CSJ) [7]. 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.

III. MULTI-MODAL SENSING WITH SMART POSTERBOARD

A) Smart posterboard

We have designed and implemented a smart posterboard, which can record poster sessions 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-inch LCD screen. Specifically, the digital posterboard is equipped with a 19-channel microphone array on the top, and attached with Kinect sensors. An outlook of the smart posterboard is given in Fig. 2. A more lightweight and portable system is realized by only using Kinect sensors, which captures audio and video signals.

B) Multi-modal sensing

Detection of participants and their multi-modal behaviors such as eye-gaze and speech using the smart posterboard is elaborated in Fig. 3.

The image processing is based on Kinect sensors to detect the persons of the audience by their face and then track their eye-gaze. The information is used in audio processing of speaker localization and voice activity detection, which are collectively referred to as speaker diarization. First, the location information of the persons can be used as

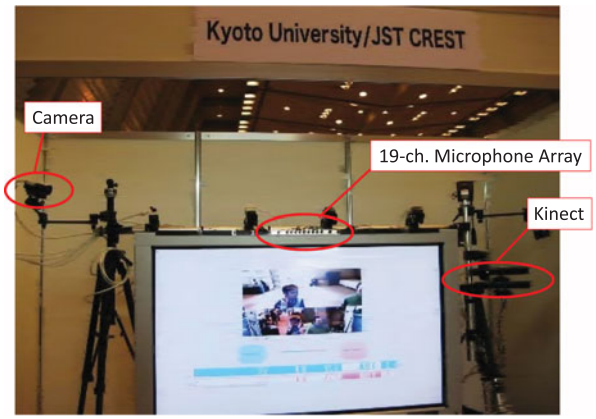

Fig. 2. Outlook of smart posterboard.

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press

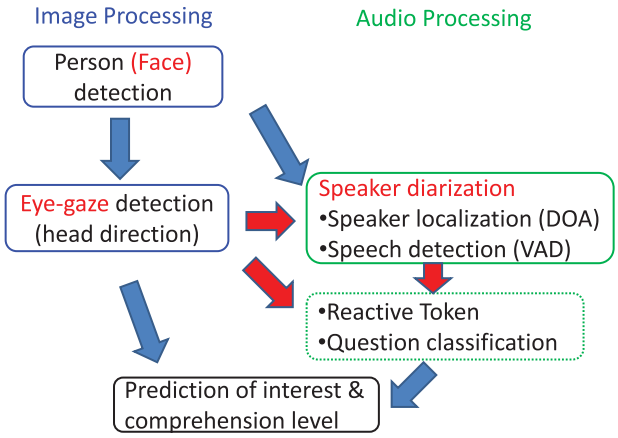

Fig. 3. Process flow of multi-modal sensing.

a constraint in the speaker localization. This is a straightforward multi-modal integration in speaker diarization [8, 9]. In this study, furthermore, we investigate the use of eye-gaze information for speaker diarization as it is shown in Section IV that eye-gaze information is useful for predicting turn-taking by the audience.

In Section VI, the eye-gaze information is also used to predict and classify speech acts by the audience, which are related with the interest and comprehension level of the audience.

C) 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. The process of the face and head orientation detection is as follows [10]:

(i) Face detection:
Haar-like features are extracted from the color and time-of-flight (ToF) images to detect the face of the participants. Multiple persons can be detected simultaneously even if they move around.

(ii) 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.

(iii) 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.

(iv) 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. In order to verify the accuracy of the head position and orientation estimated by the above method, we compared the result against the measurement by the magnetometric sensor for 16 subjects. The mean error of head position is 12.2 mm and that of the head orientation is 5.21 degrees.

IV. 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 [11–14], some studies investigated a prediction mechanism for a dialogue system to take or yield turns based on machine learning [15–18].

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 dialog 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 conversations. 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. 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 *et al.* [22] and by Bohus and Horvitz [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.

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press

In this section, four poster sessions are used. They are provided with the accurate eye-gaze annotation as explained in Section II. The ground-truth data are used in the analysis of this section. 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.

A) Analysis on eye-gaze in turn-taking

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

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 statistics are shown in Fig. 4 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 1 in relation with the turn-taking events. We can see that 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.

Joint eye-gaze events

Next, joint eye-gaze events by the presenter and the audience are defined as shown in Table 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.

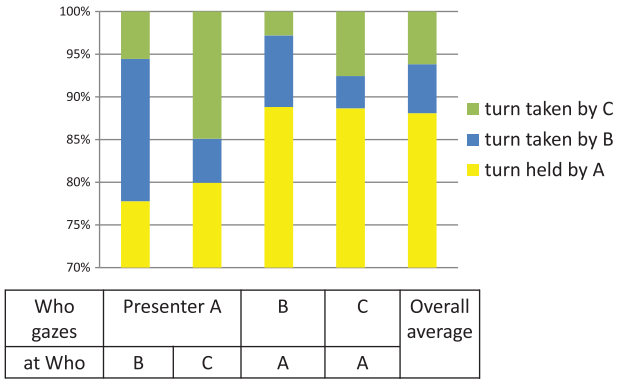

Fig. 4. Statistics of eye-gaze and its relationship with turn-taking (ratio).

Table 1. Duration of eye-gaze and its relationship with turn-taking (s).

	Turn held by Presenter A	Turn taken by audience	
		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

Table 2. Definition of joint eye-gaze events by presenter and audience.

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

Table 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 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 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.

B) 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 sub-task, we predict whether or not the turn is yielded from the presenter to (someone in) the audience, and if that happens, then we predict who in the audience

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press

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 next utterance of the current speaker (=turn-holding) or speaker change (=turn-yielding).

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.

Prediction of speaker change

For the first sub-task of prediction of speaker change, prosodic features are adopted as a baseline. Automatic speech recognition of natural conversations is very difficult, so is detection of the end of utterances based on the lexical and syntactic analysis. Therefore, prosodic features, which are extracted robustly, have been used in the previous works (e.g. [18, 22]). Specifically, Fo (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.

Eye-gaze features are defined as below:

(i) 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.
(ii) 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.
(iii) Duration of the above 1. ((I) and (i))
A maximum is taken over persons in the audience.
(iv) 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 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.

Table 4. Prediction result of speaker change.

Feature	Recall	Precision	F-measure
Prosody	0.667	0.178	0.280
Eye-gaze	0.461	0.216	0.290
Prosody + eye-gaze	**0.706**	0.209	0.319

As shown in Table 4, the prosodic features obtain a higher recall while the eye-gaze features achieve a higher 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 eye-gaze features is effective in improving both recall and precision.

Table 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

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 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 the logistic regression model; thus the prediction accuracy obtained with SVM is listed in Table 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.

V. SPEAKER DIARIZATION USING EYE-GAZE INFORMATION

Speaker diarization is a process to identify "who spoke when" in multi-party conversations. A number of diarization methods [24–26] 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 [14, 26].

Based on the finding of Section IV, we propose a novel method of speaker diarization by incorporating eye-gaze information. Although it is known that eye-gaze information can be used to predict participants' utterances, it has

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press

not been integrated in speaker diarization tasks. In the proposed multi-modal diarization method, acoustic and eye-gaze features are extracted and integrated in a stochastic manner to detect utterances.

In this study, eight poster sessions are used. For these sessions, eye-gaze information are not provided with the ground-truth annotation, but automatically captured by Kinect sensors as explained in Section III.

Since utterances by the audience are not frequent, it is difficult to detect these utterances accurately.

A) 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, 26]. An acoustic baseline method in this study is based on sound source localization using DOAs [27–30] derived from the microphone array.

To estimate a DOA, we adopt the MUltiple SIgnal Classification (MUSIC) method [31], 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 θ 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 a sound source in 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 [32].

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 ith 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 $\vec{a}_{i,t}$ of the ith participant on the time frame t consist of the MUSIC spectrum in the range:

$$\vec{a}_{i,t} = \left[M_t\left(\theta_{i,t} - \theta_B\right), \cdots, M_t\left(\theta_{i,t}\right), \cdots, M_t\left(\theta_{i,t} + \theta_B\right)\right]^T. \quad (1)$$

B) Eye-gaze features

The eye-gaze features $\vec{g}_{i,t}$ for the ith participant on the time frame t are same as those used in Section B, except that unigram and bigram of the eye-gaze objects and the joint eye-gaze events are added.

C) Integration of acoustic and eye-gaze information

The acoustic features $\vec{a}_{i,t}$ are integrated with the eye-gaze features $\vec{g}_{i,t}$ to detect the ith 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 [26]:

$$f_{i,t}(\vec{a}_{i,t}, \vec{g}_{i,t}) = \alpha\ p(v_{i,t} = 1|\vec{a}_{i,t}) + (1 - \alpha)\ p(v_{i,t} = 1|\vec{g}_{i,t}). \quad (2)$$

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 $p(v_{i,t}|\vec{a}_{i,t}, \vec{g}_{i,t})$. Compared with this joint model, the linear interpolation model has a merit that training data do not have to be aligned between the acoustic and eye-gaze features because of independency of the two discriminative models. Furthermore, the weight coefficient α 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 $p(v_{i,t}|\vec{a}_{i,t})$ [33] as

$$\alpha = \alpha_c \cdot \frac{1 - h}{1 - h_c}, \quad (3)$$

where h_c and α_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 (every 15 s).

D) Speaker diarization experiment

Logistic regression models were trained separately for the presenter and the audience by cross-validation of the eight sessions. The constrained range of the MUSIC spectrum (θ_B) is set to 10 degrees, and the MUSIC spectrum is calculated every 1 degree, thus the dimension of the acoustic features in equation (1) becomes 21. The ideal weight coefficient α_c in equation (3) is empirically set to 0.9. To evaluate performance under ambient noise, audio data were prepared by superimposing a diffusive noise recorded in a crowded place. SNR was set to 20, 15, 10, 5, and 0 dB. In real poster conversations carried out in academic conventions, SNR is expected to be about 0 to 5 dB.

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

(i) *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.

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press

(ii) *Baseline + location constraint* [9]
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, then the hypothesis is discarded.
(iii) *Acoustic-only model*
This method fixes the weight coefficient α to 1 in equation (2), and uses only the acoustic information.

For an evaluation measure, diarization error rate (DER) [34] is used in this experiment. DER consists of False Acceptance (FA), False Rejection (FR), and speaker error (SE) as below:

$$\mathrm{DER} = \frac{\#\mathrm{FA} + \#\mathrm{FR} + \#\mathrm{SE}}{\#\mathrm{S}}, \tag{4}$$

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

Table 6 lists DER 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 eye-gaze information under noisy environments expected in real poster sessions.

The weight coefficient α in equation (2) was also manually tuned where the stepping size was 0.1. In the relatively clean environment (SNR = 20 dB), the optimal weight was 0.9. On the other hand, in the noisy environments (SNR = 5 and 0 dB), the optimal weight was 0.6. These results suggest that the weight of eye-gaze features is appropriately increased in noisy environments. The average DER by the manual tuning is 12.13%, which is slightly better than the result (13.38%) by the automatic weight estimation (equation (3)). Therefore, the automatic weight estimation works reasonably according to the acoustic environment.

Moreover, we conducted an evaluation by using the measurement by the magnetometric sensor for computing the eye-gaze features. The average DER in this case is 13.31%, which is not statistically different from the result (13.38%) by the automatic estimation by the Kinect sensors.

VI. 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. In poster conversations, 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 this section, we address estimation of 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. As described in Section II, 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.

Table 6. Evaluation of speaker diarization (DER (%)).

Method		SNR (dB)						
		∞	20	15	10	5	0	average
Baseline MUSIC	[29]	15.28	20.44	28.34	42.80	64.09	87.22	43.03
Baseline + location constraint	[9]	7.76	13.66	21.18	35.49	55.27	77.81	35.20
Acoustic-only model	Equation (2) w/o $\vec{g}_{i,t}$	**6.52**	**7.60**	**9.63**	**14.20**	22.33	34.34	15.77
Multi-modal model	Equation (2)	7.35	8.55	10.73	14.23	**18.21**	**21.22**	**13.38**

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press

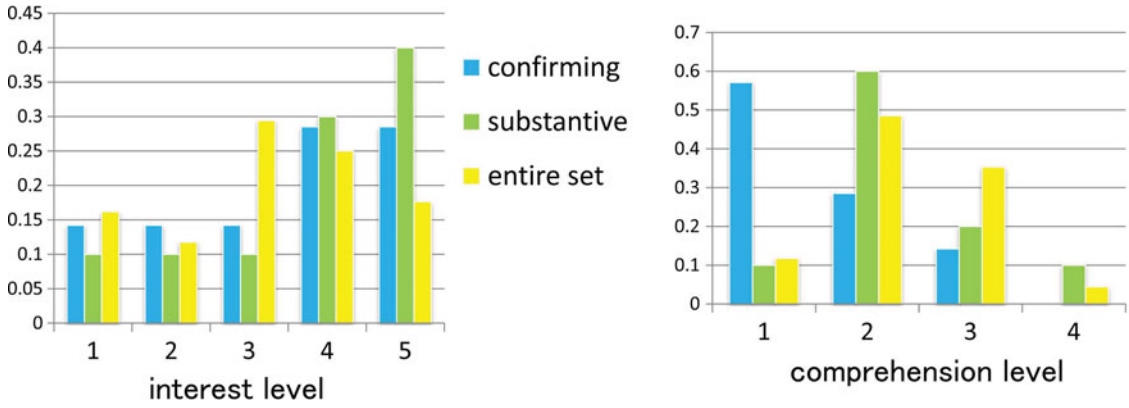

Fig. 5. Distribution of interest & comprehension level according to question type.

A) Definition of interest and comprehension level

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 work [35], 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. 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. This tendency is preliminarily analyzed with a small set in this subsection.

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"[1]. 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.

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

Relationship between question type and interest and comprehension level

In subset 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 5 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 indicate a low comprehension level (level 1&2). We also see a general tendency that occurrence of questions of either types is correlated with a higher interest level (level 4&5).

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 poster sessions or improving presentations.

B) Relationship between eye-gaze behaviors and questions

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

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 Section IV showed that eye-gaze information is related with turn-taking events; specifically, the eye-gaze by the presenter mostly controls the turn-taking.

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press

Table 7. 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

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 7. 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.

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

C) 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.

Eye-gaze at the presenter is parameterized into an occurrence count per the presenter's utterance and the duration ratio within the topic segment. A naive Bayes classifier is trained as the data size is not so large to estimate extra parameters such as weights of the features. Experimental evaluations were done by cross-validation.

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 are listed in Table 8. 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 eye-gaze features significantly improves the accuracy, but the two kinds of parameterization of the eye-gaze features (occurrence count and duration ratio) are redundant because their combination does not result in any further improvement.

Table 8. Prediction result of topic segments involving questions and/or reactive tokens.

	F-measure	Accuracy (%)
Baseline (chance rate)	0.49	49.1
Gaze occurrence	0.63	61.2
Gaze duration	0.65	57.8
Combination of both	0.63	61.2

Table 9. Identification result of confirming or substantive questions.

	Accuracy (%)
Baseline (chance rate)	51.3
Gaze occurrence	75.7
Gaze duration	67.6
Combination of both	75.7

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. In this task, the chance-rate baseline based on the prior statistic is 51.3%. The eye-gaze occurrence count achieves the best performance and combining it with the eye-gaze duration does not give an additional gain. This is explained by a large difference in its value among the two question types as shown in Table 7.

VII. 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.

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press

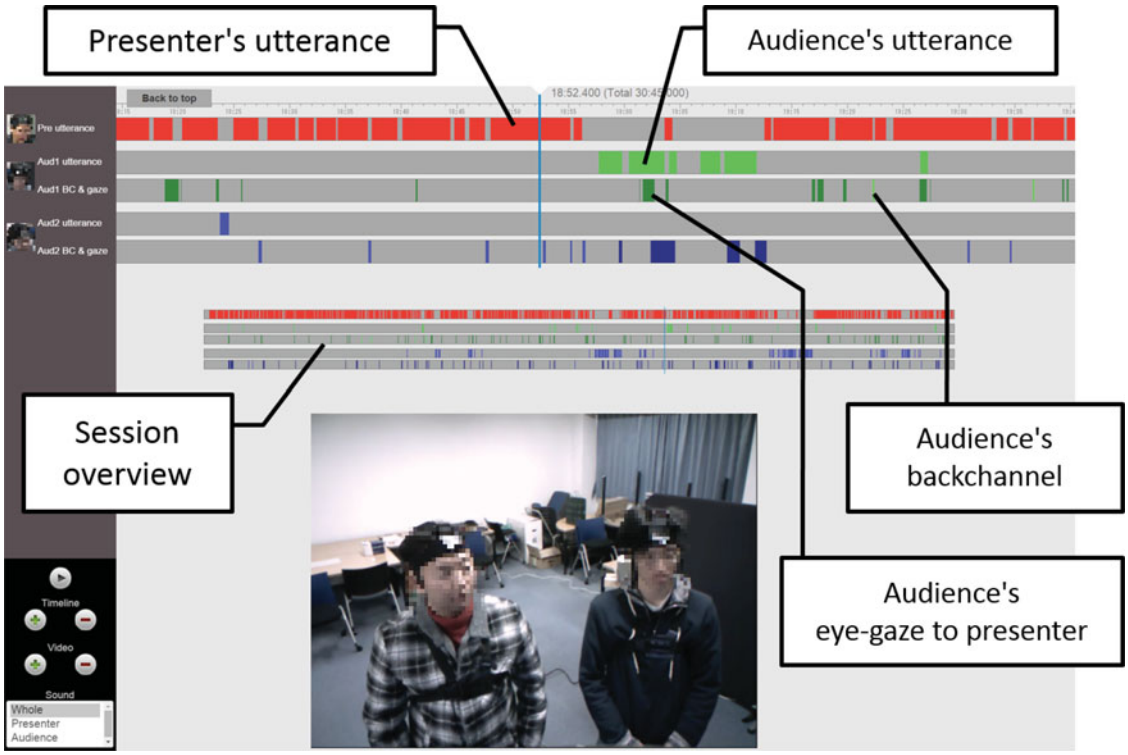

Fig. 6. Poster conversation browser.

Based on this finding, a multi-modal speaker diarization method was realized by integrating eye-gaze information with acoustic information. The stochastic multi-modal method improved the performance of speaker diarization and the effect of the 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 approached by looking into relevant 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. This scheme shows some promising results, but needs further investigations and larger-scale evaluations.

To visualize these detected events and indexes, a poster session browser has been developed, as shown in Fig. 6. Along the timeline, utterance segments of each participant are marked. We can easily access to substantial utterances from the audience such as questions and comments, which are infrequent but important in poster sessions. Eye-gaze events are also visualized so we can estimate the interaction level of the conversation. The browser will be useful for assessing the effect of the processes and further improving them.

ACKNOWLEDGEMENTS

This work was supported by JST CREST.

REFERENCES

[1] Renals, S.; Hain, T.; Bourlard, H.: Recognition and understanding of meetings: the AMI and AMIDA projects, in *Proc. IEEE Workshop Automatic Speech Recognition & Understanding*, Kyoto, 2007.

[2] Ohtsuka, K.: Conversation scene analysis. *Signal Process. Mag.*, **28**(4) (2011), 127–131.

[3] Kawahara, T.: Multi-modal sensing and analysis of poster conversations toward smart posterboard, in *Proc. SIGdial* (keynote speech), Seoul, 2012, 1–9.

[4] Kawahara, T.: Smart posterboard: multi-modal sensing and analysis of poster conversations, in *Proc. APSIPA ASC*, page (Plenary overview talk), Kaohsiung, 2013.

[5] Kawahara, T.; Sumi, K.; Chang, Z.Q.; Takanashi, K.: Detection of hot spots in poster conversations based on reactive tokens of audience, in *Proc. INTERSPEECH*, Makuhari, 2010, 3042–3045.

[6] Kawahara, T.; Setoguchi, H.; Takanashi, K.; Ishizuka, K.; Araki, S.: Multi-modal recording, analysis and indexing of poster sessions, in *Proc. INTERSPEECH*, Brisbane, 2008, 1622–1625.

[7] Maekawa, K.: Corpus of Spontaneous Japanese: its design and evaluation, in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Process. Recognit.*, Tokyo, 2003, 7–12.

[8] Nakamura, K.; Nakadai, K.; Asano, F.; Ince, G.: Intelligent sound source localization and its application to multimodal human tracking, in *Proc. IROS*, San Francisco, 2011.

[9] Wakabayashi, Y.; Inoue, K.; Yoshimoto, H.; Kawahara, T.: Speaker diarization based on audio-visual integration for smart posterboard, in *Proc. APSIPA ASC*, Siem Reap, 2014.

[10] Yoshimoto, H.; Nakamura, Y.: Cubistic representation for real-time 3D shape and pose estimation of unknown rigid object, in *Proc. ICCV Workshop*, Sydney, 2013, 522–529.

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press

[11] Ohsuga, T.; Nishida, M.; Horiuchi, Y.; Ichikawa, A.: Investigation of the relationship between turn-taking and prosodic features in spontaneous dialogue, in *Proc. INTERSPEECH*, Lisbon, 2005, 33–36.

[12] Ishi, C.T.; Ishiguro, H.; Hagita, N.: Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts, in *Proc. INTERSPEECH*, Pittsburgh, 2006, 2006–2009.

[13] Ward, N.G.; Bayyari, Y.A.: A case study in the identification of prosodic cues to turn-taking: back-channeling in Arabic, in *Proc. INTERSPEECH*, Pittsburgh, 2006, 2018–2021.

[14] Xiao, B.; Rozgic, V.; Katsamanis, A.; Baucom, B.R.; Georgiou, P.G.; Narayanan, S.: Acoustic and visual cues of turn-taking dynamics in dyadic interactions, in *Proc. INTERSPEECH*, Florence, 2011, 2441–2444.

[15] Sato, R.; Higashinaka, R.; Tamoto, M.; Nakano, M.; Aikawa, K.: Learning decision trees to determine turn-taking by spoken dialogue systems, in *Proc. ICSLP*, Denver, 2002, 861–864.

[16] Schlangen, D.: From reaction to prediction: experiments with computational models of turn-taking, in *Proc. INTERSPEECH*, Pittsburgh, 2006, 2010–2013.

[17] Raux A.; Eskenazi, M.: A finite-state turn-taking model for spoken dialog systems, in *Proc. HLT/NAACL*, Boulder, 2009.

[18] Ward, N.G.; Fuentes, O.; Vega, A.: Dialog prediction for a general model of turn-taking, in *Proc. INTERSPEECH*, Makuhari, 2010, 2662–2665.

[19] Bohus D.; Horvitz, E.: Models for multiparty engagement in open-world dialog, in *Proc. SIGdial*, London, 2009.

[20] Fujie, S.; Matsuyama, Y.; Taniyama, H.; Kobayashi, T.: Conversation robot participating in and activating a group communication, in *Proc. INTERSPEECH*, Brighton, 2009, 264–267.

[21] Laskowski, K.; Edlund, J.; Heldner, M.: A single-port non-parametric model of turn-taking in multi-party conversation, in *Proc. ICASSP*, Prague, 2011, 5600–5603.

[22] Jokinen, K.; Harada, K.; Nishida, M.; Yamamoto, S.: Turn-alignment using eye-gaze and speech in conversational interaction, in *Proc. INTERSPEECH*, Florence, 2011, 2018–2021.

[23] Kendon, A.: Some functions of gaze direction in social interaction. *Acta Psychol.*, **26** (1967), 22–63.

[24] Tranter, S.E.; Reynolds, D.A.: An overview of automatic speaker diarization systems. *IEEE Trans. ASLP*, **14**(5) (2006), 1557–1565.

[25] Reynolds, D.A.; Kenny, P.; Castaldo, F.: A study of new approaches to speaker diarization, in *Proc. INTERSPEECH*, Brighton, 2009, 1047–1050.

[26] Friedland, G. *et al.*: The ICSI RT-09 speaker diarization system. *IEEE Trans. ASLP*, **20**(2) (2012), 371–381.

[27] Macho, D. *et al.*: Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus, in *Proc. ICME*, Amsterdam, 2005, 876–879.

[28] Anguera, X.; Wooters, C.; Hernando, J.: Acoustic beamforming for speaker diarization of meetings. *IEEE Trans. ASLP*, **15**(7) (2007), 2011–2022.

[29] Araki, S.; Fujimoto, M.; Ishizuka, K.; Sawada, H.; Makino, S.: A DOA based speaker diarization system for real meetings, in *Proc. HSCMA*, Trento, 2008, 29–32.

[30] Ishiguro, K.; Yamada, T.; Araki, S.; Nakatani, T.; Sawada, H.: Probabilistic speaker diarization with bag-of-words representations of speaker angle information. *IEEE Trans. ASLP*, **20**(2) (2012), 447–460.

[31] Schmidt, R.: Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.*, **34**(3) (1986), 276–280.

[32] Yamamoto, K.; Asano, F.; Yamada, T.; Kitawaki, N.: Detection of overlapping speech in meetings using support vector machines and support vector regression. *IEICE Trans.*, **E89-A**(8) (2006), 2158–2165.

[33] Misra, H.; Bourlard, H.; Tyagi, V.: New entropy based combination rules in hmm/ann multi-stream asr, in *Proc. ICASSP*, Hong Kong, vol. **2**, 2003, 741–744.

[34] Fiscus, J.G.; Ajot, J.; Michel, M.; Garofolo, J.S.: The Rich Transcription 2006 Spring Meeting Recognition Evaluation. Bethesda, 2006.

[35] Kawahara, T.; Chang, Z.Q.; Takanashi, K.: Analysis on prosodic features of Japanese reactive tokens in poster conversations, in *Proc. Int. Conf. on Speech Prosody*, Chicago, 2010.

Tatsuya Kawahara received his B.E. degree in 1987, M.E. in 1989, and Ph.D. in 1995, all in Information Science, from Kyoto University, Kyoto, Japan. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT. He has published more than 300 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several speech-related projects in Japan including speech recognition software Julius and the automatic transcription system for the Japanese Parliament (Diet). Kawahara received the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (MEXT) in 2012. From 2003 to 2006, he was a member of IEEE SPS Speech Technical Committee. He was a general chair of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2007). He also served as a Tutorial Chair of INTERSPEECH 2010 and a Local Arrangement Chair of ICASSP 2012. He is an editorial board member of Elsevier Journal of Computer Speech and Language, APSIPA Transactions on Signal and Information Processing, and IEEE/ACM Transactions on Audio, Speech, and Language Processing. He is VP-Publications (BoG member) of APSIPA and a senior member of IEEE.

Takuma Iwatate graduated from Kyoto University, Japan with M.S. degree in Informatics in 2012.

Koji Inoue received his B.E. degree in 2013 from Kurume National College of Technology, Japan, and M.S. degree in informatics in 2015 from Kyoto University, Japan, respectively. He is currently pursuing a Ph.D. degree at Graduate School of Informatics, Kyoto University, and has been a Research Fellow of the Japan Society for the Promotion of Science (JSPS) since 2015. His research interests include multi-modal interaction analysis, multi-modal signal processing, and spoken dialogue systems. He is a student member of IEEE.

Soichiro Hayashi graduated from Kyoto University, Japan with M.S. degree in Informatics in 2013.

Hiromasa Yoshimoto received his B.Sc. and M.Sc. in Engineering from Kyushu University, Japan, in 2000 and 2002, respectively. From 2005 through 2015, he was a Researcher at Kyoto University. Currently, he works as a Researcher at The University of Tokyo. His research interests include computer vision and visual tracking. He is a member of IPSJ, IEICE, and HIS.

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press

Katsuya Takanashi received his B.A. degree in Faculty of Letters in 1995, M.A. of Human and Environmental Studies in Graduate School of Human and Environmental Studies in 1997, and Ph.D. in Graduate School of Informatics in 2014, from Kyoto University, Kyoto, Japan. From 2000 to 2005 he was a Researcher at National Institute of Information and Communications Technology, Kyoto, Japan. Currently, he is a Researcher in Graduate School of Informatics, Kyoto University. He has been studying multi-modal and multi-party human-human and human-robot interaction from both cognitive scientific and sociological perspectives. He has been a representative of a project on field studies in situated multi-party conversation and also a member of several other projects on the modeling of multi-modal interaction in Japan.

https://doi.org/10.1017/ATSIP.2016.2 Published online by Cambridge University Press