Enhanced speaker diarization with detection of backchannels using eye-gaze information in poster conversations

Koji Inoue[1], *Yukoh Wakabayashi*[2], *Hiromasa Yoshimoto*[3], *Katsuya Takanashi*[3], *and Tatsuya Kawahara*[1,3]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]Graduate School of Information Science and Engineering, Ritsumeikan University, Japan
[3]Academic Center for Computing and Media Studies, Kyoto University, Japan

Abstract

We propose multi-modal speaker diarization using acoustic and eye-gaze information in poster conversations. Eye-gaze information plays an important role in turn-taking, thus it is useful for predicting speech activity. In this paper, a variety of eye-gaze features are elaborated and combined with the acoustic information by the multi-modal integration model. Moreover, we introduce another model to detect backchannels, which involve different eye-gaze behaviors. This enhances the diarization result by filtering meaningful utterances such as questions and comments. Experimental evaluations in real poster sessions demonstrate that eye-gaze information contributes to improvement of diarization accuracy under noisy environments, and its weight is automatically determined according to the Signal-to-Noise Ratio (SNR).

Index Terms: speaker diarization, backchannel, multi-modal, eye-gaze, poster conversation

1. Introduction

Recent advances in sensing technologies allow us to conduct multi-modal analyses and processing of multi-party conversations. In the AMI / AMIDA [1] and VACE [2] projects, multi-party conversations such as meetings were recorded using multiple microphones and cameras. Analyses involved the multi-modal data including not only verbal information but also non-verbal channels, such as backchannels, nodding, and eye-gaze behaviors [3, 4].

We have been collecting multi-modal data on conversations in poster sessions (= poster conversations), where a presenter makes an interactive presentation to a small audience. This conversation form is commonly conducted in academic conventions including InterSpeech conferences. To analyze poster conversations, a recording environment, called "smart posterboard", equipped with multi-modal sensing devices such as a microphone array and cameras has been developed [5]. Based on the collected multi-modal data, we have investigated turn-taking behaviors [6] and interest / comprehension levels of the audience [7].

This study addresses speaker diarization in poster conversations. Speaker diarization is to identify "who spoke when" in multi-party conversations. A number of diarization methods [8, 9, 10] 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 [8] and distant talking [11]. To solve the problem, some studies tried to incorporate multi-modal information such as motion and gesture [12, 10, 13].

We propose a multi-modal diarization method which integrates eye-gaze information with acoustic information. The eye-gaze behavior plays an important role in turn-taking in multi-party conversations [14, 15]. For example, it is often observed that eye-gaze directions of participants tend to intersect each other when they change the conversational turn. Although it is known that eye-gaze information can be used to predict participants' utterances [16, 6, 17], eye-gaze information has not been integrated in speaker diarization tasks. In our previous work [18], we proposed a multi-modal scheme which integrates eye-gaze information with acoustic information, and conducted preliminary experiments. The proposed method extracts acoustic and eye-gaze features, which are integrated in a stochastic manner. In this paper, we conduct additional data collection, and make substantial improvements in the eye-gaze features and the multi-modal integration model. Furthermore, the diarization results are enhanced by detecting audience's backchannels to efficiently review the conversational content in poster sessions. Detection of backchannels is also realized by using the same multi-modal scheme. Backchannels are frequently observed in poster conversations and involve different eye-gaze behaviors since they indicate that the listener does not take a turn. Thus, a different model is trained for their detection. 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 the interaction level of the conversation.

2. Multi-modal corpus of poster conversations

The smart posterboard system consists of a 19-channel microphone array, Kinect sensors, and HD cameras, which are attached to a large LCD (Figure 1). With this setting, eight poster sessions were recorded. In each session, one presenter made a poster presentation on his/her academic research, and there was an audience of two persons, standing in front of the poster and listening to the presentation. The duration of each session was 20 to 30 minutes. Speaker diarization and detection of backchannels are conducted with the sensors (the microphone array and Kinect) attached to the posterboard, so the participants do not have to wear any devices. For the ground truth annotation of the data used in this work, however, speech data were also recorded with a wireless headset microphone, and participants' head locations and orientations were captured by magnetometric sensors.

Table 1 summarizes the total utterance duration in the recorded sessions. While the presenter (*pre*) holds a majority

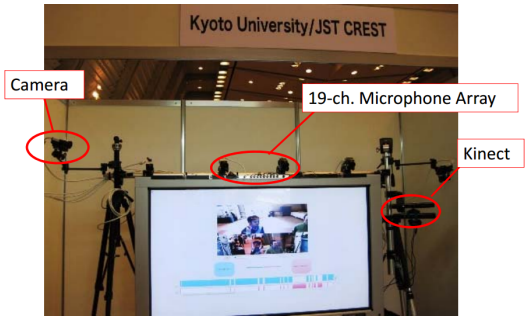

Figure 1: Outlook of smart posterboard

Table 1: Total utterance duration [sec.] (backchannels)

session ID	*pre*	*aud1*	*aud2*	total
140206-01	1,251	19 (11)	227 (111)	1,497
140206-02	1,406	283 (138)	164 (15)	1,853
140206-03	1,333	328 (160)	170 (86)	1,831
140206-04	1,495	129 (57)	102 (35)	1,726
140207-01	1,343	164 (48)	123 (21)	1,630
140207-02	1,229	134 (52)	117 (26)	1,480
140207-03	1,205	106 (41)	267 (79)	1,578
140207-04	1,208	216 (113)	135 (81)	1,559
total	10,470	2,684 (1,074)		13,154

of the turns, utterances by the audience (*aud1* and *aud2*) are not frequent, which means it would be difficult to detect these utterances accurately. The audience's backchannels account for about 40 percent of their utterance duration.

3. Multi-modal speaker diarization

This section describes the proposed speaker diarization method which integrates acoustic and eye-gaze information.

3.1. Baseline acoustic method

Conventional speaker diarization methods have used Mel-Frequency Cepstral Coefficients (MFCCs) [8, 10] and Directions Of Arrival (DOA) of sound sources [19, 20, 21, 22] as acoustic features. 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 [23, 24], 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, 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 an SVM using the eigenvalue distribution of a spatial correlation matrix [25].

3.2. Multi-modal method integrating eye-gaze information

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 combines the two probabilities for the frame-wise decision. The process is conducted independently on every time frame t and for each participant i.

3.2.1. *Acoustic features*

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:

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

3.2.2. *Eye-gaze features*

The eye-gaze direction used in this study is approximated by the head orientation estimated from RGB and depth images captured by the Kinect sensors. The head orientations are tracked by a particle filter [26]. The eye-gaze object is determined by the head-orientation vector and the location of the objects. In this work, the object is limited to the poster and the other participants.

The eye-gaze features are designed based on the eye-gaze objects of the i-th participant and a conversational partner[1]. The eye-gaze feature vector $\boldsymbol{g}_{i,t}$ consists of the followings:

1. Eye-gaze object
   This feature represents which object the i-th participant looks at. For the presenter, (*P*) poster or (*I*) audience; For (anybody in) the audience, (*p*) poster or (*i*) presenter.
2. Joint eye-gaze event: "*Ii*", "*Ip*", "*Pi*", "*Pp*"
   This feature represents combination of the eye-gaze objects by the i-th participant and the conversational partner. For example, "*Ii*" and "*Pp*" correspond to mutual gaze and joint attention, respectively.
3. Maximum duration of each state of the above 1.
4. Maximum duration of each state of the above 2.
5. Uni-gram and Bi-gram of the above 1.
   This feature represents the transition of the eye-gaze objects.
6. Uni-gram and Bi-gram of the conversational partner's eye-gaze objects

The first and the second features are calculated for the time frame t, and the others are measured within the preceding period (the preceding C ms).

3.2.3. *Multi-modal integration model*

We integrate the acoustic features $\boldsymbol{a}_{i,t}$ with the eye-gaze features $\boldsymbol{g}_{i,t}$ to detect the i-th participant's speech activity $v_{i,t}$ in the time frame t. The speech activity $v_{i,t}$ is binary: speaking ($v_{i,t} = 1$) or not-speaking ($v_{i,t} = 0$). We adopt a linear interpolation of the probabilities independently computed by the two feature sets[2] [10]:

$$\begin{aligned} f_{i,t}(\boldsymbol{a}_{i,t}, \boldsymbol{g}_{i,t}) \;=\; & \alpha\, p(v_{i,t} = 1|\boldsymbol{a}_{i,t}) \\ & + (1-\alpha)\, p(v_{i,t} = 1|\boldsymbol{g}_{i,t}) \end{aligned} \quad (1)$$

Here, $\alpha \in [0, 1]$ is a weight coefficient. Each probability is computed by a logistic regression model. It is also possible

[1] The conversational partner of the presenter is the audience, and the conversational partner of the audience is the presenter.

[2] In the previous work [18], we used a generative model to compute $p(\boldsymbol{a}_{i,t}|v_{i,t})$, but it is difficult to estimate the weight coefficient α because of a large difference of the dynamic ranges.

to combine the two feature sets in the feature domain and directly compute a posterior probability $p(v_{i,t}|\boldsymbol{a}_{i,t}, \boldsymbol{g}_{i,t})$. Compared with this joint model, the linear interpolation model has a merit that training data does not need 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 $p(v_{i,t}|\boldsymbol{a}_{i,t})$ [27, 28] as

$$\alpha = \alpha_c \cdot \frac{1-h}{1-h_c}\ , \tag{2}$$

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 every 15 seconds by using the preceding 15-second data.

4. Detection of backchannels

The diarization result includes backchannels and also falsely accepted noise especially for audience's utterances. We introduce a post-processing model 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, for example, MFCCs and Gaussian Mixture Model (GMM) are used to classify backchannels and other acoustic events [29]. On the other hand, many studies have been conducted to predict appropriate timing of backchannels [30, 31, 32, 33, 34].

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, we train a different model using the eye-gaze behaviors to predict backchannels. Here, we adapt the multi-modal scheme formalized in the previous section. The eye-gaze features and the multi-modal integration model are the 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-ΔMFCCs) [29]
3. Power (and ΔPower)
4. Regression coefficients of fundamental frequency (F0) and power at the end of the preceding utterance [32]

Logistic regression models are trained to predict three events: backchannels, utterances other than backchannels, and noise. For each utterance segment as a result of 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.

5. Experimental evaluations

The proposed methods are evaluated using the corpus of poster conversations mentioned in Section 2.

5.1. Setup

Logistic regression models were trained separately for the presenter and the audience by cross-validation of the eight sessions. To evaluate one session, the other seven sessions were used for training. In the proposed multi-modal method, the SVM to determine the number of sound sources and the entropy h_c of a clean acoustic environment were estimated with the training data. The constrained range of the MUSIC spectrum of the acoustic features was 10 degrees ($\theta_B = 10^\circ$). This setting was intended to prevent overlapping between the participants. Since the MUSIC spectrum was calculated every 1 degree, the dimension of the acoustic features was 21. The scope to calculate the duration and Uni- and Bi-gram of the eye-gaze features was 1,000 ms ($C = 1,000$). Frame rates per second of the acoustic and eye-gaze features are 62.5 and 29.5, respectively. Consequently, an interpolation using the nearest samples was done to the eye-gaze features. The ideal weight coefficient α_c (Eq. (2)) in a clean environment was set to: 0.9 for speaker diarization and 0.8 for detection of backchannels.

To evaluate performance under ambient noise, we prepared audio data by superimposing a diffusive noise recorded in a crowded place on the audio signals. SNRs were set to 20, 15, 10, 5 and 0 dB. In real poster sessions carried out in academic conventions, the SNRs are expected to be around 0 to 5 dB.

5.2. Speaker diarization result

We compared the proposed multi-modal method with other methods listed below:

1. *baseline MUSIC* [21]
   This method conducts peak tracking of the MUSIC spectrum and GMM-based clustering in the angle domain. Each cluster corresponds to a participant. This method does not use any cue from visual information.
2. *baseline + location constraint* [35]
   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 α to 1 in Eq. (1), and uses only the acoustic information.

For an evaluation measure, Diarization Error Rate (DER) [36] was used in this experiment. DER consists of False Acceptance (FA), False Rejection (FR), and Speaker Error (SE) as below:

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

where $\#$S is the number of speech frames in the reference data. This metric does not evaluate $\pm$ 250 ms collars of the reference utterance segments. We evaluated the minimum DER by varying the threshold for Eq. (1) after smoothing (hangover) the detected utterance segments.

Table 2 lists DERs for each SNR. The two baseline methods (*baseline MUSIC* and *baseline + location constraint*) showed lower accuracy than the stochastic methods (*acoustic-only model* and *multi-modal model*) because the baseline methods 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 achieved higher performance under the noisy environments (SNR $= 5, 0$ dB). Thus, we can see the effect of the

Table 2: Evaluation of speaker diarization (DER [%])

method		SNR [dB]						
		∞	20	15	10	5	0	average
baseline MUSIC	[21]	16.94	23.14	31.66	47.92	67.03	88.80	45.92
baseline + location constraint	[35]	8.34	14.45	22.31	36.09	55.80	78.05	35.84
acoustic-only model	eq. (1) w/o $\boldsymbol{g}_{i,t}$	**6.16**	**7.28**	**9.36**	14.20	22.94	35.89	15.97
multi-modal model	eq. (1)	6.27	7.81	9.96	**13.69**	**18.18**	**21.61**	**12.92**

Table 3: Evaluation of audience's speech detection (EER [%])

method		SNR [dB]						
		∞	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. (1) w/o $\boldsymbol{g}_{i,t}$	**12.14**	**13.98**	15.47	**18.19**	23.34	30.20	18.89
multi-modal model	eq. (1)	12.23	14.11	**15.42**	18.29	**23.07**	**29.72**	**18.80**

eye-gaze information under noisy environments expected in real poster sessions.

We also manually tuned the weight coefficient α in Eq. (1) 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 $= 5, 0$ dB), the optimal weights were 0.6 or 0.5. These results suggest that the weight of eye-gaze features is adequately increased in noisy environments. The average DER by the manual tuning was 11.78%, which was slightly better than the result (12.92%) by the automatic weight estimation (Eq. (2)). Therefore, the automatic weight estimation works reasonably according to the acoustic environment.

5.3. Effect of backchannel detection

The diarization result was post-processed by another model for elimination of backchannels and noise. In this experiment, backchannel segments in the reference data were regarded as non-speech events. We compared the following methods, which were applied after the proposed multi-modal diarization method (last row of Table 2).

1. *thresholding with utterance duration*
   A threshold in this method is the duration of each utterance segment since the duration of backchannels is usually shorter than others. This corresponds to using only the first feature listed in Section 4.

2. *acoustic-only model*
   This method uses the acoustic features listed in Section 4 to detect backchannels and noise.

3. *multi-modal model*
   This method also uses the eye-gaze features in addition to the acoustic features.

The thresholds in the detection of backchannels were empirically determined to: 800 ms for *thresholding with utterance duration* and 0.8 for the posterior probabilities in *acoustic-only model* and *multi-modal model*.

Here, we focused 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:

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

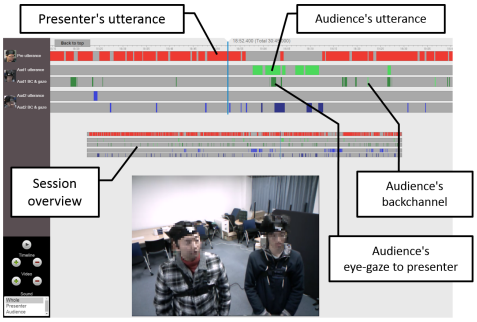

Figure 2: Poster session browser

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

Table 3 lists EERs for each SNR. Compared to the case without post-processing (*no post-processing*), the proposed multi-modal model significantly reduced EERs. This shows the effectiveness of elimination of backchannels and noise after speaker diarization. The simple thresholding method (*thresholding with utterance duration*) reduced EERs in noisy conditions, but degraded 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).

6. Conclusions

We have proposed a multi-modal speaker diarization method which integrates eye-gaze information with acoustic information. Moreover, the diarization result is 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 is confirmed under noisy environments, which are expected in real poster sessions.

For an application to visualize the diarization results, we developed a poster session browser (Figure 2), which can be executed on Web browsers. This application enables us to efficiently review the meaningful utterances such as questions and comments, together with the interaction level from the detected backchannels and eye-gaze events in the poster conversation.

7. Acknowledgements

This work was supported by JSPS KAK-ENHI Grant Number 15J07337 and JST CREST / ERATO.

8. References

[1] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 28–39.

[2] L. Chen, R. T. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, Y. Xiong, D. McNeill, R. Tuttle, and T. Huang, "VACE multimodal meeting corpus," in *Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 40–51.

[3] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[4] K. Otsuka, "Conversation scene analysis," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 127–131, 2011.

[5] T. Kawahara, "Smart posterboard: Multi-modal sensing and analysis of poster conversations," in *Proc. APSIPA ASC*, 2013, pp. 1–5.

[6] T. Kawahara, T. Iwatate, and K. Takanashi, "Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations," in *Proc. INTERSPEECH*, 2012, pp. 727–730.

[7] T. Kawahara, S. Hayashi, and K. Takanashi, "Estimation of interest and comprehension level of audience through multi-modal behaviors in poster conversations," in *Proc. INTERSPEECH*, 2013, pp. 25–29.

[8] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. ASLP*, vol. 14, no. 5, pp. 1557–1565, 2006.

[9] D. A. Reynolds, P. Kenny, and F. Castaldo, "A study of new approaches to speaker diarization," in *Proc. INTERSPEECH*, 2009, pp. 1047–1050.

[10] G. Friedland, A. Janin, D. Imseng, X. A. Miro, L. Gottlieb, M. Huijbregts, M. T. Knox, and O. Vinyals, "The ICSI RT-09 speaker diarization system," *IEEE Trans. ASLP*, vol. 20, no. 2, pp. 371–381, 2012.

[11] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: mixing acoustic features and inter-channel time differences," in *Proc. INTERSPEECH*, 2006.

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

[13] B. G. Gebre, P. Wittenburg, S. Drude, M. Huijbregts, and T. Heskes, "Speaker diarization using gesture and speech," in *Proc. INTERSPEECH*, 2014, pp. 582–586.

[14] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta psychologica*, vol. 26, no. 1, pp. 22–63, 1967.

[15] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of personality and social psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[16] K. Jokinen, K. Harada, M. Nishida, and S. Yamamoto, "Turn-alignment using eye-gaze and speech in conversational interaction," in *Proc. INTERSPEECH*, 2010, pp. 2018–2021.

[17] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings," in *Proc. ICASSP*, 2014, pp. 694–698.

[18] K. Inoue, Y. Wakabayashi, H. Yoshimoto, and T. Kawahara, "Speaker diarization using eye-gaze information in multi-party conversations," in *Proc. INTERSPEECH*, 2014, pp. 562–566.

[19] D. Macho, J. Padrell, A. Abad, C. Nadeu, J. Hernando, J. McDonough, M. Wolfel, U. Klee, M. Omologo, A. Brutti, P. Svaizer, G. Potamianos, and S. M. Chu, "Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus," in *Proc. ICME*, 2005, pp. 876–879.

[20] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. ASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

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

[22] K. Ishiguro, T. Yamada, S. Araki, T. Nakatani, and H. Sawada, "Probabilistic speaker diarization with bag-of-words representations of speaker angle information," *IEEE Trans. ASLP*, vol. 20, no. 2, pp. 447–460, 2012.

[23] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[24] Y. Huang, T. Otsuka, and H. G. Okuno, "A speaker diarization system with robust speaker localization and voice activity detection," in *Contemporary Challenges and Solutions in Applied Artificial Intelligence*. Springer, 2013, pp. 77–82.

[25] K. Yamamoto, F. Asano, T. Yamada, and N. Kitawaki, "Detection of overlapping speech in meetings using support vector machines and support vector regression," *IEICE Trans. Fundamentals*, vol. 89, no. 8, pp. 2158–2165, 2006.

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

[27] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in hmm/ann multi-stream asr," in *Proc. ICASSP*, vol. 2, 2003, pp. 741–744.

[28] K. Iwano, T. Matsuo, and S. Furui, "A study on unsupervised stream-weight estimation for multimodal speech recognition," in *IPSJ SIG Report*, ser. SLP-76-24, 2009, pp. 1–6.

[29] T. Kawahara, K. Sumi, Z. Chang, and K. Takanashi, "Detection of hot spots in poster conversations based on reactive tokens of audience," in *Proc. INTERSPEECH*, 2010, pp. 3042–3045.

[30] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs," *Language and speech*, vol. 41, no. 3-4, pp. 295–321, 1998.

[31] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[32] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa, "Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems," *Journal of JSAI*, vol. 20, no. 3, pp. 220–228, 2005.

[33] L. P. Morency, I. D. Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 70–84, 2010.

[34] D. Ozkan and L. P. Morency, "Modeling wisdom of crowds using latent mixture of discriminative experts," in *Proc. ACL*, 2011, pp. 335–340.

[35] Y. Wakabayashi, K. Inoue, H. Yoshimoto, and T. Kawahara, "Speaker diarization based on audio-visual integration for smart posterboard," in *Proc. APSIPA ASC*, 2014.

[36] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, *The rich transcription 2006 spring meeting recognition evaluation*. Springer, 2006.