AUDIO-VISUAL CONVERSATION ANALYSIS BY SMART POSTERBOARD AND HUMANOID ROBOT

Tatsuya Kawahara, Koji Inoue, Divesh Lala, Katsuya Takanashi

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

This paper addresses audio-visual signal processing for conversation analysis, which involves multi-modal behavior detection and mental-state recognition. We have investigated prediction of turn-taking by the audience in a poster session from their multi-modal behaviors, and found out that the eye-gaze provides an important cue compared with head nodding and verbal backchannels. This finding has been applied to audio-visual speaker diarization by combining eye-gaze information. We are now investigating engagement recognition in human-robot interaction based on the same scheme. Robust and real-time detection of laughing, backchannels and nodding is realized based on LSTM-CTC. We introduce a latent "character" model to cope with the subjectivity and variations of engagement annotations. Experimental evaluations demonstrate that (1) the latent character model is effective, (2) automatic behavior detection is robust and does not degrade the engagement recognition accuracy, and (3) the eye-gaze is the most important feature among others.

Index Terms— Audio-visual signal processing, conversation analysis, behavior analysis, engagement, human-robot interaction

1. INTRODUCTION

In human-human interaction, we exhibit our mental states and attitudes via several behaviors even without speaking. The behaviors while listening include eye-gaze, head nodding, and verbal backchannels. These suggest attentiveness, engagement and interest. Occasionally, strong reactions such as laughing and assessment tokens (e.g. "*wow*") are observed [1]. Detection of these multi-modal behaviors and recognition of mental states are an important skill in human communication, and thus required for artificial intelligence. For example, an agent in a digital signage should detect the audience's attention, and should stop talking when the audience does not attend any more (even if they are present) [2]. Or it should take questions if the audience is interested and has something to say. This skill of conversational analysis or "mood sensing" is critical for a humanoid robot with a human-like appearance and interaction functionality.

We conducted the "smart posterboard" project, which consists of multi-modal sensing and analysis of conversations in poster sessions [3, 4, 5]. Poster sessions are a norm in many conferences and open-lab events because they allow for flexible and interactive presentations. The audience can show feedback to the presentation in real time, and the presenters are expected to take questions even during the presentation and, if necessary, switch the content and explanation according to them. Thus, we set up an interesting problem whether we can predict the audience's turn-taking based on the behaviors during the attendance [5]. We also empirically know that the audience's questions and comments suggest their interest in the presentation. Thus, the conversational analysis of a recorded session in an offline mode is useful for estimating the interest level of the audience in the poster session [6]. We also investigated audio-visual speaker diarization that detects when the audience makes utterances by combining the eye-gaze information with the audio information [7]. In this paper, we highlight the importance of the eye-gaze behaviors among others.

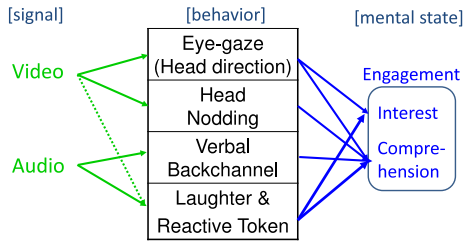

Fig. 1. Scheme of audio-visual conversation analysis

Currently, we are conducting a project on symbiotic human-robot interaction with a goal of an autonomous android robot who behaves and interacts just like a human [8, 9, 10]. When we consider a social role of the robot such as a receptionist and a lab guide, who plays a role of poster presenter, the first step is to sense the human attitude toward the interaction, particularly the engagement level. Therefore, we investigate engagement recognition based on the audio-visual processing of the user's behaviors.

Fig. 1 depicts the scheme of audio-visual conversation analysis, which has been applied to two domains addressed in this paper. In the domain of smart posterboard, we tried to predict turn-taking or detect speaking activity of the audience, which are objectively observed and can be regarded as an approximation of their engagement in the poster session. In the domain of human-robot interaction, we investigate engagement recognition. As the annotation of engagement is subjective, we introduce a latent "character" model to cope with a variety of annotators. We have designed dedicated audio-visual sensing systems and implemented signal processing modules to detect multi-modal behaviors. LSTM-CTC has realized robust detection of behaviors based on the event-wise optimization in training. We will also investigate which behaviors have impact on these prediction and recognition tasks.

2. CONVERSATION SCENE ANALYSIS OF POSTER SESSIONS BY SMART POSTERBOARD

We have designed and implemented a smart posterboard, which can record poster sessions and sense human behaviors. The posterboard, which is actually a 65-inch LCD screen, is equipped with a 19-

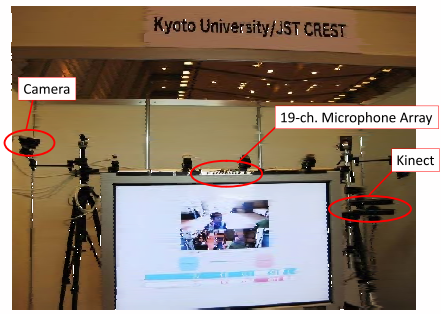

Fig. 2. Outlook of smart posterboard

channel microphone array on the top and attached with Kinect sensors. An outlook of the smart posterboard is shown in Fig. 2. A more lightweight and portable system is realized by only using Kinect sensors. A set of high-resolution cameras were also used for corpus recording.

We have recorded a number of poster sessions using this device. In each session, one presenter prepared a poster on his/her own 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 20-30 minutes.

2.1. Prediction of Turn-taking by Audience in Poster Sessions

Turn-taking in conversations is a natural behavior by humans, but it is still challenging for spoken dialogue systems and conversational robots. Recently, a number of studies have been conducted to model and implement natural turn-taking functions [11, 12, 13, 14, 15], but a majority of them are still focused on dyadic conversations between two persons or between a user and a system. There are a few studies that deal with meetings and conversations by more than two persons [16, 17].

Conversations in poster sessions are different from those settings, in that presenters hold most of turns and thus the amount of utterances is very unbalanced. However, the segments of the audience's questions and comments are more informative and should not be missed. We also presume that the audience signals the willingness to take a turn via multi-modal behaviors. Therefore, we set up a problem to predict turn-taking by the audience using multi-modal behaviors.

The prediction is 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). Since there are multiple persons in the audience, turn-taking or turn-yielding is counted by either person of the audience.

Prosodic features of the presenter's utterance were adopted as a baseline based on the previous work [17]. Specifically, F0 (mean, max, min and range) and power (mean and max) of the presenter's utterance was computed prior to the prediction point. Each feature was normalized by the speaker by taking the z-score.

In this study, we particularly focus on the effect of multi-modal behaviors of the audience. We have incorporated nodding and backchannels as well as eye-gaze behaviors. We simply counted head nodding from visual information and verbal backchannels from audio information. Eye-gaze features were defined by the eye-gaze object (poster or audience or presenter) and the joint eye-gaze event, which was a combination of the eye-gaze object of the presenter and that of the audience, and the duration of these [5].

Table 1. Prediction result of turn-taking by audience in poster sessions

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

Prediction experiments were conducted by using four sessions in a cross-validation manner. In this experiment, the ground-truth annotations of backchannels and eye-gaze information were used. The results with SVM classifiers are listed in Table 1. Here, recall, precision and F-measure were computed for 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 turn-taking, considering the nature of the task and application scenarios.

As shown in Table 1, the baseline prosodic features obtained a higher recall while the eye-gaze features achieved a higher precision and F-measure. Combination of the eye-gaze features with the prosodic features was effective for improving both recall and precision. On the other hand, the backchannel feature got the lowest performance, and its combination with other features resulted in degradation of the performance. This result demonstrated that the eye-gaze behavior provides a strong cue in turn-taking while backchannels do not necessarily show strong engagement.

2.2. Audio-Visual Speaker Diarization in Poster Sessions

Since the eye-gaze information provides a cue for turn-taking as shown in the previous subsection, it is expected to be useful for detecting speaking activity. Therefore, we implemented a multi-modal speaker diarization by incorporating eye-gaze information. Speaker diarization is a process to identify "who spoke when" in multi-party conversations, and it has been investigated based on audio information. In real multi-party conversations, the diarization performance is degraded by adversary acoustic conditions such as background noise and distant talking.

An acoustic baseline method was based on sound source localization using DOAs (Direction Of Arrivals) derived from the microphone array. To estimate a DOA, we adopted the MUltiple SIgnal Classification (MUSIC) method [18], which can detect multiple DOAs simultaneously. The MUSIC spectrum $m_t(\theta)$ was calculated based on the orthogonal property between an input audio signal and a noise subspace. Here 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. We can also use the participant's head location tracked by the Kinect sensors. The possible location of the i-th participant is constrained within a certain range ($\pm\theta_B$) from the detected location $\theta_{i,t}$. Thus, we define the acoustic feature $\mathbf{a}_{i,t}$ of the i-th participant at time frame t with the MUSIC spectrum in the range:

$$\mathbf{a}_{i,t} = [m_t(\theta_{i,t} - \theta_B), \cdots, m_t(\theta_{i,t}), \cdots, m_t(\theta_{i,t} + \theta_B)]^T \quad (1)$$

Table 2. Evaluation of audio-visual speaker diarization (DER [%])

method		SNR [dB]						
		∞	20	15	10	5	0	average
acoustic-only model	eq. (3) w/o $\mathbf{g}_{i,t}$	**6.52**	**7.60**	**9.63**	**14.20**	22.33	34.34	15.77
feature-level combination	eq. (2)	6.95	7.91	9.85	15.12	26.43	43.66	18.32
likelihood-level combination	eq. (3)	7.35	8.55	10.73	14.23	**18.21**	**21.22**	**13.38**

Then, we incorporate eye-gaze information extracted from visual information. The eye-gaze feature $\mathbf{g}_{i,t}$ for the i-th participant at time frame t is essentially same as those used in the previous subsection, except that the features are computed for every time frame using the preceding frames. The acoustic feature $\mathbf{a}_{i,t}$ and the eye-gaze feature $\mathbf{g}_{i,t}$ are integrated to detect the i-th participant's speech activity $v_{i,t}$ at time frame t. Note that the speech activity $v_{i,t}$ is binary: speaking ($v_{i,t} = 1$) or not-speaking ($v_{i,t} = 0$).

In this study, we make a comparison of two integration methods: feature-level combination and likelihood-level combination. The feature-level combination trains a single classifier which takes a combined input of the acoustic feature and the eye-gaze feature.

$$f_{i,t}(\mathbf{a}_{i,t}, \mathbf{g}_{i,t}) = p(v_{i,t} = 1|\mathbf{a}_{i,t}, \mathbf{g}_{i,t}) \quad (2)$$

The likelihood-level combination conducts a linear interpolation of the likelihoods independently computed by the two feature sets.

$$f_{i,t}(\mathbf{a}_{i,t}, \mathbf{g}_{i,t}) = \alpha \cdot p(v_{i,t} = 1|\mathbf{a}_{i,t}) + (1-\alpha) \cdot p(v_{i,t} = 1|\mathbf{g}_{i,t}) \quad (3)$$

Here $\alpha \in [0, 1]$ is a weight coefficient. Each likelihood is computed by a logistic regression model to take a value $\in [0, 1]$. Compared with the feature-level combination, the likelihood-level combination has a merit that training data does not have to be aligned between the acoustic and eye-gaze features. Furthermore, the weight coefficient α can be appropriately determined according to the acoustic environments such as Signal-to-Noise Ratio (SNR). Here, it is estimated using an entropy h of the acoustic posterior probability $p(v_{i,t}|\mathbf{a}_{i,t})$ [19].

In this experiment, eight poster sessions were used in a cross-validation manner. Eye-gaze information was automatically captured by Kinect sensors. Logistic regression models were trained respectively for the presenter and the audience. To evaluate performance under ambient noise, audio data was 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 around 0 to 5 dB.

Table 2 lists Diarization Error Rate (DER) for each SNR. Compared with the acoustic-only model, the audio-visual likelihood-level combination achieved higher performance under noisy environments (SNR = 5, 0 dB). Thus, we confirm the effect of eye-gaze information under noisy environments expected in real poster sessions. On the other hand, the feature-level combination did not work well because the weight of the two features were fixed during the training and cannot be adjusted according to SNR.

For reference, we tuned the weight coefficient α in Eq. (3) manually with the stepping size of 0.1. In the relatively clean environment (SNR = 20 dB), the optimal weight was 0.9, but it was 0.6 in the noisy environments (SNR = 5 and 0 dB). These results suggest that the weight of eye-gaze features must be and could be increased appropriately in noisy environments. The average DER by the manual tuning is 12.13%, which is only slightly better than the result (13.38%) by the automatic weight estimation. Therefore, the automatic weight estimation works reasonably according to the acoustic environment.

3. ENGAGEMENT RECOGNITION IN HUMAN-ROBOT INTERACTION

In the previous section, we investigated prediction of turn-taking by the audience in poster sessions. Since turn-taking suggests engagement in the session, this scheme can be naturally applied to detection of the engagement level. Recognition of user engagement is particularly required for agents and robots interacting with humans [2, 16, 20]. The agent or robot can keep talking the current topic if the user is engaged in the conversation, but otherwise should stop talking or change the topic. In this study, we investigate engagement recognition in human-robot interaction based on multi-modal behaviors, as depicted in Fig. 1.

We are developing an autonomous android ERICA who looks, behaves and interacts just like a human. She is designed to play a social role such as a receptionist and a lab guide with natural spoken dialogue as well as non-verbal behaviors such as gazing and nodding. With this human-like android, we expect users to exhibit behaviors just as in human-human interactions, in which a variety of multi-modal behaviors signal engagement.

We have collected a number of conversation sessions with ERICA in a Wizard-of-Oz (WOZ) setting. The dialogue was recorded with directed microphones, a 16-channel microphone array, RGB cameras, and Kinect v2 sensors. An outlook of a session and the recording environment are shown in Fig. 3.

We had five annotators to label the engagement level of the user for each conversation session (12 annotators and 20 sessions in total). Here, we focus on the listening mode of the users and multimodal behaviors, so instructed the annotators to label for each turn of the robot if the user's engagement level is regarded as high based on some behavior cues. Behaviors that suggest high engagement include facial expressions, verbal backchannels, head nodding, eye-gaze, laughing and body movements. However, mapping from these behaviors to the engagement level, or how to interpret them, may be subjective and different for each annotator. Instead of simply taking a mean or a majority of the labels given by multiple annotations, we introduce a Bayesian model, in which the engagement labels are generated via a latent character of the annotator.

In this model, given a behavior pattern b_k, the engagement level e is labeled by an annotator a_i via a latent character variable c_j, where e is binary (high (e=1) or not (e=0)) and c_j is discrete.

$$p(e|b_k, a_i) = \sum_{j}^{J} p(e|c_j, b_k) p(c_j|a_i) \quad (4)$$

The behavior patterns b_k are defined by observed combinations of individual behaviors (i.e. laughing, backchannels, nodding and eye-gaze) to consider their co-occurrence effects. Then, the two kinds of probabilities are estimated via collapsed Gibbs sampling, which samples each character alternatively and iteratively from the conditional probability distributions. We also tried different sizes of the latent characters c_j, and found J=4 provides the best performance.

Fig. 3. Dialogue with an android ERICA in WOZ setting

This model deals with variations of annotators with robust estimation, and provides engagement-level prediction per a given annotator. In human-robot interaction, we designate a character of ERICA.

We also implement automatic detection of laughing, backchannels, nodding and eye-gaze, though detection of the facial expressions is yet to be done. Detection of laughing and backchannels is performed with bidirectional-LSTM and the CTC (Connectionist Temporal Classification) criterion using the audio information [21]. The CTC allows for event-wise optimization in the training of the detection model without precise frame-wise labels and will integrate it with automatic speech recognition in a unified framework. In this experiment, we used audio recorded with a directed microphone, and computed standard log-Mel filterbank features (10msec shift). Then, LSTM of five hidden layers with 256 nodes per each layer was trained with the CTC criterion using 71 dialogue sessions, which contains 3931 backchannels and 1003 laughing samples. An evaluation on the 20 test sessions shows that precision and recall of backchannels were 0.780 and 0.865, and those of laughing were 0.772 and 0.496.

Detection of nodding and eye-gaze is done using the visual information, in particular, the head position and pose captured by the Kinect v2 sensor. For nodding detection, we use a feature set of instantaneous speed of the yaw, roll, and pitch of the head together with the average speed, velocity, acceleration and the range of the head pitch over the previous 500msec. It is fed to another LSTM of a single hidden layer with 16 nodes, which outputs a posterior probability of nodding at every 10msec frame. In a cross-validation using the 20 sessions, which contains 855 nodding events, the recall and precision were 0.608 and 0.763, respectively. The eye-gaze behavior is modeled by a logistic regression model to take a value of 1 when the user is gazing the robot longer than a threshold. Eye-gaze toward the robot is detected when the distance between the head-orientation vector and the location of the robot's head is smaller than a threshold. This detection is conducted every 10msec.

As these models are designed to generate a probability $p(b_k|\mathbf{x})$ of detecting a behavior b_k given an audio-visual observation $\mathbf{x}$, where $\mathbf{x}$ is the feature set mentioned above, the engagement level recognition is formulated as below:

$$p(e|\mathbf{x}, a_i) = \sum_{j}^{J} \sum_{k}^{K} p(e|c_j, b_k) p(c_j|a_i) p(b_k|\mathbf{x}) \qquad (5)$$

In this implementation, combination of the multi-modal behaviors is done in the definition of the behavior patterns b_k and their summation in the above formula.

The overall system is realized according to the scheme depicted in Fig. 1, which consists of the two steps of behavior detection and engagement recognition. While the behavior detection modules are trained based on objective annotations, the engagement recognition model takes into account subjective annotations.

Table 3. Engagement recognition accuracy (%) in human-robot interaction

behaviors	manual annotation	automatic detection
all (J=1)	0.674	0.663
all (J=4)	0.711	0.700
w/o backchannel	0.699	0.684
w/o laughing	0.684	0.689
w/o nodding	0.700	0.699
w/o eye-gaze	0.681	0.669

Engagement recognition experiments were conducted using the 20 sessions in a cross-validation manner. Table 3 lists the results in terms of recognition accuracy when the behaviors are manually given (Eq. 4) and when they are automatically detected from audio-visual information (Eq. 5). The table first shows the effect of the latent character model. With the character size of 4 (J=4), the recognition accuracy is much improved from the model without considering the latent characters (J=1), which is comparable to a simple logistic regression model. There is only a little degradation with automatic detection, and the result demonstrates the applicability in real-world setting. The table also lists the performance by removing each behavior one by one. The results show that the eye-gaze behavior is the most critical. Laughing also makes some contribution [22], but backchannels and nodding do not have a significant impact to overall recognition. It is suggested that backchannels and nodding express some reaction, but not engagement. The results are in accordance with the findings in the posterboard domain addressed in the previous section.

All behavior detection modules are implemented to allow for real-time engagement recognition. This enables the robot to adaptively switch the action according to the user's engagement level.

4. CONCLUDING REMARKS

We have investigated audio-visual signal processing for conversation analysis, which consists of multi-modal behavior detection and mental-state recognition, in two application domains. One is the smart posterboard, which senses the audience's behaviors in poster sessions for intelligent media archiving. It is shown that the eye-gaze provides effective features for turn-taking prediction and speaker diarization in noisy conditions. This finding can be extended to an intelligent interaction system to conduct poster presentations.

Therefore, we are working on the other application of a humanoid robot who behaves and interacts just like a human. Recognition of user engagement is modeled and implemented via detection of multi-modal behaviors such as backchannels, nodding, laughing and eye-gaze. The experimental evaluations demonstrate that the contribution of the eye-gaze is the most significant. We also demonstrate the feasibility of real-time engagement recognition based on audio-visual signal processing with reasonable performance.

The results in the two domains confirm that the proper eye contact is important in conversations, especially for expressing positive feedback, and thus must be realized by humanoid robots in both recognition and generation.

Acknowledgment: This work was supported by JST CREST program and ERATO Ishiguro Symbiotic Human-Robot Interaction program (Grant Number JPMJER1401).

5. REFERENCES

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

[2] Y.I.Nakano and R.Ishii, "Estimating user's egnagement from eye-gaze behaviors in human-agent interaction," in *Proc. IUI*, 2010.

[3] T.Kawahara, "Multi-modal sensing and analysis of poster conversations toward smart posterboard," in *Proc. SIGdial Meeting Discourse & Dialogue*, 2012, pp. 1–9 (keynote speech).

[4] T.Kawahara, "Smart posterboard: Multi-modal sensing and analysis of poster conversations," in *Proc. APSIPA ASC*, 2013, p. (plenary overview talk).

[5] T.Kawahara, T.Iwatate, K.Inoue, S.Hayashi, H.Yoshimoto, and K.Takanashi, "Multi-modal sensing and analysis of poster conversations with smart posterboard," *APSIPA Trans. Signal & Information Process.*, vol. 5, no. e2, pp. 1–12, 2016.

[6] T.Kawahara, S.Hayashi, and K.Takanashi, "Estimation of interest and comprehension level of audience through multimodal behaviors in poster conversations," in *Proc. INTERSPEECH*, 2013, pp. 1882–1885.

[7] K.Inoue, Y.Wakabayashi, H.Yoshimoto, K.Takanashi, and T.Kawahara, "Enhanced speaker diarization with detection of backchannels using eye-gaze information in poster conversations," in *Proc. INTERSPEECH*, 2015, pp. 3086–3090.

[8] D.F.Glas, T.Minato, C.T.Ishi, T.Kawahara, and H.Ishiguro, "ERICA: The ERATO Intelligent Conversational Android," in *Proc. RO-MAN*, 2016, pp. 22–29.

[9] K.Inoue, P.Milhorat, D.Lala, T.Zhao, and T.Kawahara, "Talking with ERICA, an autonomous android," in *Proc. SIGdial Meeting Discourse & Dialogue*, 2016, vol. Demo. Paper, pp. 212–215.

[10] P.Milhorat, D.Lala, K.Inoue, Z.Tianyu, M.Ishida, K.Takanashi, S.Nakamura, and T.Kawahara, "A conversational dialogue manager for the humanoid robot ERICA," in *Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS)*, 2017.

[11] G.Skantze, A.Hjalmarsson, and C.Oertel, "Turn-taking, feedback and joint attention in situated human-robot interaction," *Speech Communication*, vol. 65, pp. 50–66, 2014.

[12] D.Kim, C.Breslin, P.Tsiakoulis, M.Gasic, M.Henderson, and S.Young, "Inverse reinforcement learning for micro-turn management," in *Proc. InterSpeech*, 2014, pp. 328–332.

[13] J.Kane, I.Yanushevskaya, C.Looze, B.Vaughan, and A.N.Chasaide, "Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions," in *Proc. InterSpeech*, 2014, pp. 333–337.

[14] A.Gravano, P.Brusso, and S.Benus, "Who do you think will speak next? perception of turn-taking cues in Slovak and Argentine Spanish," in *Proc. InterSpeech*, 2016, pp. 1265–1269.

[15] R.Masumura, T.Asami, H.Masataki, R.Ishii, and R.Higashinaka, "Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks," in *Proc. InterSpeech*, 2017, pp. 1661–1665.

[16] D.Bohus and E.Horvitz, "Models for multiparty engagement in open-world dialog," in *Proc. SIGdial*, 2009.

[17] K.Jokinen, K.Harada, M.Nishida, and S.Yamamoto, "Turn-alignment using eye-gaze and speech in conversational interaction," in *Proc. InterSpeech*, 2011, pp. 2018–2021.

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

[19] H.Bourlard H.Misra and V.Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Proc. IEEE-ICASSP*, 2003, vol. 1, pp. 741–744.

[20] Z.Yu, L.Nicolich-Henkin, A.W.Black, and A.Rudnicky, "A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement," in *Proc. SIGdial*, 2016.

[21] H.Inaguma, K.Inoue, M.Mimura, and T.Kawahara, "Social signal detection in spontaneous dialogue using bidirectional LSTM-CTC," in *Proc. INTERSPEECH*, 2017, pp. 1691–1695.

[22] B.B.Turker, Z.Bucinca, E.Erzin, Y.Yemez, and M.Sezgin, "Analysis of engagement and user experience with a laughter responsive social robot," in *Proc. InterSpeech*, 2017, pp. 844–848.