Leader Identification Using Multimodal Information in Multi-party Conversations

Tsukasa Shiota†, Kouki Honda†, Kazutaka Shimada, and Takeshi Saitoh
Department of Artificial Intelligence, Kyushu Institute of Technology
Iizuka, Fukuoka, Japan
{t_shiota, k_honda, shimada}@pluto.ai.kyutech.ac.jp, saitoh@cse.kyutech.ac.jp
†*The 1st and 2nd authors contributed equally to this research.*

***Abstract*—It is one of the important tasks to predict a participant's role in a multi-party conversation. Many previous studies utilized only verbal or non-verbal features to construct models for the role recognition task. In this paper, we propose a model that combines verbal and non-verbal features for leader identification. We add non-verbal features and construct our prediction model with utterance, pose, facial, and prosodic features. In our experiments, we compare our model with a baseline model that is based on only utterance features. The results show the effectiveness of our multimodal approach. In addition, we improve the performance of the baseline model to add some new utterance features.**

***Index Terms*—speaker role identification, spoken language, multi-party conversation understanding, multimodal analysis**

I. Introduction

As the field of natural language processing has been growing, many researchers have put much effort into making a computer do assessments, such as Automated Essay Scoring [1]. These assessment activities do not usually have the correct answers due to a variety of responses, so it is a very difficult task to estimate the scores. Meanwhile, some studies focused on spoken discussions and attempted to assess the quality of the participant's behavior in discussions. For example, Okada et al. [2] have proposed a model to estimate communication skills for each participant in group discussions. It is also a challenging task and our objective is close to the latter study.

To realize an automatic assessment system for spoken discussions, identifying the roles of the participants is required. In detail, there could be some types of people in a discussion: a person that puts effort into eliciting other's opinions and making a consensus in the discussion, and a person that comes up with many ideas and accelerates the discussion. If the system evaluates the two persons on the same dimension, e.g. creativity, the former person would receive a low score although s/he contributed to the discussion differently. In other words, the system should choose assessment dimensions on the basis of the role of each participant in the discussion. For this reason, it is an important task to identify the roles of the participants in conversations. Some researchers tackled various role identification tasks (see Section II), but most of the research considered only verbal or non-verbal characteristics of the participants. People usually express what they think or feel with not only language but also gesture and voice tone in face-to-face conversations. Therefore, utilizing multimodal information is an integral element for the role identification task.

In this paper, we tackle a role identification task in multi-party conversations. We propose an identification model that classifies each participant in a discussion into a leader or non-leader with multimodal features. For the model, we introduce both utterance features reported in the previous work [3] and non-verbal features proposed by our study. As the non-verbal features, we design pose, facial, and prosodic features. In addition, we incorporate additional utterance features on the basis of knowledge from Rienks et al [4].

Our contributions are as follows:

- We generate a high accuracy model based on multimodal features for the leader identification task. Our model outperformed a baseline model with only verbal features by [3].
- We compare the effectiveness of each modality. We examine the best combination of features by a feature ablation test for leader identification.
- We add some new utterance features. These features contributed to the improvement in the performance of the verbal feature-based method.

II. Related Work

Many studies on role recognition have been conducted. Benne and Sheats [5] have defined functional roles that divide participants in working groups into 3 classes including 28 specific roles. To apply the roles in [5] to face-to-face multi-party meetings, Zancaro et al. [6] have proposed the Functional Role Coding Scheme. In the scheme, they afresh defined 10 roles based on the functional roles and they categorize the roles into 2 types of role categories: Task Area (Orienteer, Giver, Seeker, Recorder, and Follower) and Socio-Emotional Area (Attacker, Gate-keeper, Protagonist, Supporter, and Neutral). Assuming that the roles dynamically continue to change during a meeting, they annotated the roles on each short duration of voice activity and then verified the accuracy of automatic detection. Huang et al. [7] have redefined the functional roles and predicted them annotated on utterances. These studies focused on predicting the functions of each utterance. On the other hand, our study aims to identify the overall role of each participant throughout a discussion.

Wilson et al. [8] have employed linguistic and prosodic features as well as speaker's subjectivity for their role recognition model. Their experiment results showed that the model improved its performance when combining all the features. Considering this fact, we expect that performance of leader identification will be improved if we employ both utterance features and non-verbal features adequately.

Shiota et al. [3] have attempted to distinguish leaders from non-leaders in multi-party conversations. They designed features that relate to utterance and constructed a leader identification model. Sanchez-Cortes et al. [9] have focused on estimating leadership rather than identifying specific roles. They did not tackle the functional role recognition but Emergent Leader (EL) detection. They constructed the ELEA corpus and detected the participants that naturally behaved as a leader in the discussions. Beyan et al. [10] have predicted the most and the least EL-like participants in the ELEA corpus. Our study is close to them, but those studies only introduced verbal or non-verbal features. We apply multimodal features to our task.

III. Target Data

In this paper, we utilize the Kyutech corpus, which is a freely available Japanese multi-party conversational corpus constructed by Yamamura et al. [11]. In each meeting, four participants work on a decision-making task. The participants play the roles of managers in a virtual shopping mall and decide which restaurant is the most suitable for their shopping mall from three candidates instead of a closed restaurant. The corpus contains 9 meetings and each of them consists of utterance transcriptions and video data. In addition, Yamamura et al. [11], [12] annotated a topic tag and dialogue act tag to each Japanese long utterance unit [13]. Yamamura et al. defined 28 topic tags, such as "Close" (the topic about a closed restaurant), and 22 dialogue act tags based on ISO24617-2 [14], such as "Question" (the act to obtain an answer or some information).

In the corpus, any participants are not assigned the roles. However, the corpus contains a questionnaire to every participant about the satisfaction of the decision. In detail, it contains a question, "Who do you think which participant led the meeting?" Shiota et al. [3] regarded the participants that are voted the most by the question as the ground truth in their task. They also verified the reliability of the subjective evaluation data with third parties. They reported that it was not a problem that the subjective evaluation data, namely the majority vote from the questionnaire, was regarded as the ground truth. In this paper, we also use their data set in the track of their report.

IV. Proposed Method

In this section, we propose our method to predict leaders in meetings using multimodal features. Figure 1 shows the outline of our proposed model. We handle the method in [3] as our baseline (the bottom side of the figure). They used utterance features for the task. In this section, we explain the baseline features first.

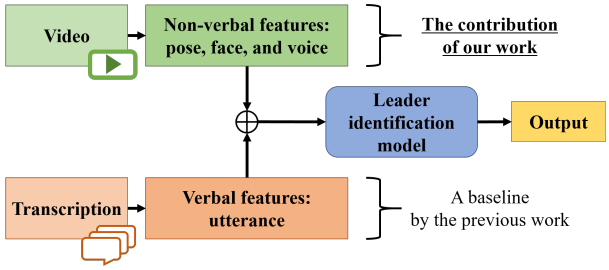

Fig. 1. The outline of the proposed method

Our contribution is to incorporate non-verbal features in the baseline. The contribution is described in Section IV-B. Our features consist of the pose, facial, and prosodic features extracted from videos in the corpus.

A. Baseline Features

The baseline features by Shiota et al. [3] can be extracted from the utterance transcriptions in the corpus. We briefly explain all the utterance features below.

B_1 The ratios of the number of repetition utterances by him/herself and by the other participants
In a meeting, participants repeat previous utterances in order to elicit more information and to show that they understand what is being talked about. Therefore, the 2 types of repetition have some clues for leader identification. For the repetition extraction, each utterance is converted into a BoW vector. Then, the cosine similarity is computed between a target utterance and each utterance in the next 10 utterances. If the similarity exceeds a threshold, the sentence in the next 10 sentences is regarded as the repetition. After that, the number of the repetition utterances by him/herself and by the other participants are counted up respectively, and the ratios are computed for each participant.

B_2 The coverage of the topics in a meeting
Leaders tend to speak on most topics in a meeting to control how the meeting goes. Therefore, the number of topics in which each participant utter is counted up, and the coverage for each participant is computed.

B_3 The ratio of the number of "Meeting" tag
One of the leader's role is to guide a discussion in the right direction. In other words, leaders are likely to speak when the group is talking about the proceedings and decisions of the discussion. Therefore, the number of utterances with the "Meeting" tag (the topic about the proceedings and final decision) is counted up, and the ratio for each participant is computed.

B_4 The ratios of the number of dialogue acts
Leaders are more likely to elicit opinions and assessments from the other participants while non-leaders often tend to show their statements. It means

Fig. 2. Estimated information by OpenPose (left) and OpenFace (right)

that leader's acts and non-leader's ones in a meeting could differ. We count up the number of utterances with 5 dialogue acts (Question, Answer, Inform, Suggest, PositiveFeedback), and the 5 ratios for each participant are computed.

B_5 **The ratios of the number of utterances in the whole and each quarter**
Leaders have to speak at an appropriate timing to make a meeting go smoothly. As a result, they would utter more than non-leaders in a meeting. Therefore, the number of utterances in each quarter and the whole meeting are counted up, and the ratios for each participant are computed.

B_6 **The average length and time of utterances**
Leaders sometimes accelerate or prohibit other participants' statements or activities. As a result, utterances of leaders might differ from non-leaders' ones, so the averages of length (the number of characters) and time of utterances for each participant are computed.

B_7 **The ratio of the number of utterances after a silence**
A meeting does not always go smoothly because the participants might have to make a tough decision, etc.; consequently, no one utters. A leader is obligated to break through it, so the silence has a clue to predict the leader in the participants. Therefore, the number of utterances after a silence is counted up, and the ratio for each participant is computed. In this article, a silence is defined as an interval that there are not any utterances more than 10 seconds.

B. Proposed Non-verbal Features

Multimodal information is a promising element for improvement in the identification model. Thus, we introduce pose, facial, and prosodic features to our model.

To extract the pose of all the participants in a meeting from video data, we utilize OpenPose [15] that estimates body keypoints in each frame of a video. The left side of Figure 2 shows an example of estimated body information with OpenPose. We compute the pose features introduced below for 12 keypoints (nose, left eye, right eye, left ear, right ear, neck, left shoulder, right shoulder, left elbow, right elbow, left wrist, and right wrist.)

P_1 **The standard deviations of x and y coordinates**
Leaders look around the other participants or move their bodies during meetings to engage participants. In other words, the movement of the body keypoints of leaders and non-leaders would be different. Therefore, we compute standard deviations of x and y coordinates of the 12 keypoints for each participant.

P_2 **The mean values and standard deviations of moving amount of x and y coordinates**
Even though we employ the standard deviations of x and y coordinates of the keypoints, modulation of the movement of each point cannot be captured. Therefore, we also compute mean values and standard deviations of the moving amount of x and y coordinates for each participant.

Leaders have not only their characteristics on body movement but also their ones on their face, such as eye gaze and head pose. Although OpenPose can estimate x and y coordinates of facial information, such as eyes, OpenFace [16] is a more superior technology to capture facial information. The right side of Figure 2 shows an example of estimated facial information with OpenFace. OpenFace can estimate facial landmarks, eye gaze, and head pose in each frame of a video. We compute the facial features below with the results of the estimation.

F_1 **The standard deviations of x and y coordinates of the facial landmarks**

F_2 **The mean values and standard deviations of moving amount of x and y coordinates of the facial landmarks**
Leaders are more likely to look around to make the others comfortable during a meeting; in other words, their faces often move. Therefore, we compute standard deviations and the mean values of them for each participant.

F_3 **The standard deviations of the eye gaze**
Same as F_1 and F_2, leaders' eye gazes also vary widely. Therefore, We compute standard deviations of the left eye gaze, right eye gaze, and the average of the eyes gaze.

F_4 **The standard deviations of the head position and its rotation angle**
We should also consider the head position, so we compute standard deviations of the head position and its rotation angle estimated by OpenFace.

We also extract prosodic features, such as voice pitch and voice volume, to capture the differences between leaders and non-leaders in terms of acoustic information. We acquire each participant's voice from the video by using timestamps in the utterance transcriptions in the target corpus. For each voice segmentation, we extract the prosodic features below with LibROSA [17].

V_1 **MFCC**
V_2 **RMS**
V_3 **Fundamental frequency**

V_4 Spectrum center of gravity
V_5 Spectral contrast

In this paper, the maximum value, minimum value, greatest difference, and standard deviation of the prosodic features for all the utterances are computed. We employ the mean values for those values for our leader identification model.

V. Identification Experiment

In this section, we verify the performance of our leader identification model with multimodal features. We compare our model with a simple baseline based on verbal features by [3]. We evaluate the effectiveness and contribution of each feature by an ablation experiment.

A. Experimental Setting

We conducted a leader identification experiment as a binary classification task; namely, a model classifies a participant into a leader or non-leader. As mentioned in Section III, the gold standard is based on the results of the questionnaires. According to [3] that constructed the data set, eight leaders were determined for eight meetings by the voting. For the rest, namely one meeting, two participants received the same number of votes. In the data set construction, they regarded the two participants as the leaders in the meeting. To summarize, the leader identification task in the data set is to detect 10 leaders in 36 participants (9 meetings x 4 participants).

We evaluated all of the models with conversational-level leave-one-out cross-validation. It means that a model is trained with 8 meetings, and the rest of 1 meeting is used for evaluation. We report the average of the 9 iterations. We employed a decision tree model called CART [18] for our classification model (seed value = 0)[1].

B. Result

Table I shows the result of the leader identification experiment. P, R, and F1 on the 2nd line are the abbreviations of precision, recall, and F-measure respectively. B, P, F, V, and ALL on the 1st column denote a set of utterance features (B_1 to B_7), a set of pose features (P_1 and P_2), a set of facial features (F_1 to F_4), a set of prosodic features (V_1 to V_5), and a set of all of the proposed features (ALL) respectively. Scores in bold indicate that they are higher than the score of the baseline (B), and those with underline are the maximum values in each column.

For the models with unimodal features (B, F, P, and V), the results of the models with only non-verbal features (F, P, and V) were lower than the baseline model (B), namely the model with utterance features. The model with utterance features was a strong baseline in terms of the unimodal setting. On the other hand, only non-verbal features cannot identify leaders well.

[1]We evaluated several machine learning methods, such as SVMs. The CART algorithm was stable and high performance. Therefore, we discuss the experiment by the result of the CART. In addition, our final goal is to realize an automatic behavior assessment system. Thus, visible and understandable approaches are more preferred because the model can show the reason why the person is regarded as the role.

TABLE I
Results of Leader Identification

Feature set	Leader			Non-leader		
	P	R	F1	P	R	F1
B	0.667	0.611	0.638	0.861	0.926	0.892
P	0.204	0.278	0.235	0.713	0.630	0.669
F	0.055	0.111	0.074	0.574	0.537	0.555
V	0.111	0.111	0.111	0.667	0.741	0.702
B+P	0.537	0.667	0.595	**0.889**	0.889	0.889
B+F	**0.704**	**0.722**	**0.713**	**0.907**	0.926	**0.917**
B+V	0.500	**0.667**	0.574	**0.880**	0.852	0.866
P+F	0.426	0.556	0.482	0.852	0.852	0.852
P+V	0.056	0.111	0.074	0.648	0.667	0.657
F+V	0.083	0.333	0.133	0.435	0.519	0.473
B+P+F	**0.778**	**0.833**	**0.805**	**0.935**	0.926	**0.931**
B+P+V	**0.722**	**0.722**	**0.722**	**0.907**	**0.963**	**0.934**
B+F+V	0.315	0.389	0.348	0.815	0.852	0.833
P+F+V	0.093	0.222	0.131	0.657	0.630	0.643
ALL (ours)	**0.889**	**0.833**	**0.860**	**0.935**	**1.000**	**0.967**

Next, we compared the baseline model (B) with the multimodal models. Some multimodal models using utterance features (B+F, B+P+F, B+P+V, ALL) improved their leader identification accuracy. Furthermore, our model (ALL), namely the model with all the features, produced the best performance in all the criteria. In contrast, some models that do not use the utterance features (P+F, P+V, F+V, P+F+V) did not perform well. For these results, it is suggested that multimodal models including the utterance features are much more effective than the baseline model (B).

In summary, the unimodal models with only non-verbal features we introduced were not effective for leader identification. Some multimodal models outperformed the baseline for leader identification; however, utterance features were essential.

C. Model Analysis

A benefit of employing a decision tree model is to be able to understand how a model learns the characteristics of leaders and non-leaders. Therefore, we visualize some trained models and analyze which features are selected in detail. Figure 3 and Figure 4 show examples of the trained decision trees by the only baseline's feature (B) and by all the features (ALL) respectively. Each node contains three types of information: (1) a feature and its threshold that are used to classify participants in the node, (2) the number of participants in the node, and (3) the distribution of the two roles in the node (left-hand number in the square bracket indicates the number of non-leaders and the other does the number of leaders). The tree in Figure 3 used 3 features listed below.

- The ratio of the number of utterances that request more information from the others (QU)
- The ratio of the number of repetition utterances by him/herself (self_repetition)
- The average time duration of utterances (ave_time)

As can be seen in Figure 3 and 4, QU appeared on both trees. It denotes that the feature plays an important role in

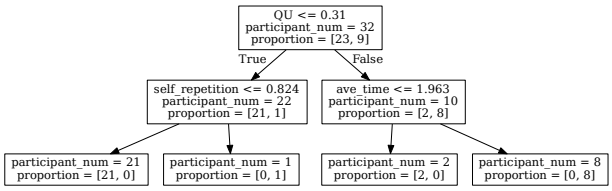

Fig. 3. A tree generated by the baseline

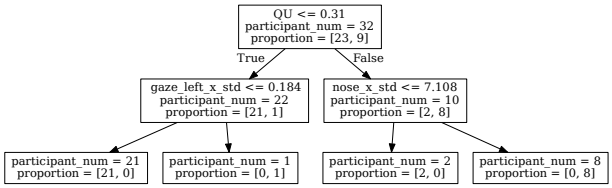

Fig. 4. A tree generated by the ALL model

leader identification. Moreover, the average time duration of utterances also classifies leaders well. From the tree in Figure 3, it is interpreted that leaders are more likely to ask questions to the other participants, and their speech tends to become long.

The tree in Figure 4 used two non-verbal features listed below other than QU.

- The standard deviation of left eye gaze in the horizontal direction (gaze_left_x_std)
- The standard deviation of the x coordinate of nose (nose_x_std)

The features represent pieces of the face parts, and those parts do not move independently. In other words, it is reasonable to think that the features indicate the movement of the head. In addition, both features are related to horizontal movement (x-axis), so it is more important to measure the horizontal movement of the head for leader identification. From the tree in Figure 4, it is interpreted that leaders are more likely to ask questions, and they tend to horizontally move their head, e.g., to look around the other participants.

To sum up, the result implies that leaders were more likely to elicit information or opinions from others, to speak longer than the others, and to look around the other participants. It is indicated that the features that represent the semantic aspect of utterances and the movement of the face are especially effective for leader identification.

VI. Comparison with improved baseline

According to the results we reported in Section V-B, features from utterances possess an important clue for leader identification. In this section, we strengthen the baseline, which is the model with utterance features (B), and compare the improved baseline with our multimodal model (ALL).

A. Additional Utterance Features

We generate a more powerful baseline to reveal the effectiveness of our multimodal method. The baseline model by [3] does not consider some characteristics of interaction, such as turn-taking. Therefore, we introduced some new utterance features regarding interaction on the basis of the paper by Rienks et al. [4].

B_1^* The ratio of the number of topic initialization
Leaders tend to monitor and control meeting topics in order to make the discussions go smoothly. We count up the number of the utterances that change a topic in the meeting and the ratio for each participant is computed.

B_2^* The ratio of the number of turn-takings
A turn in [4] was defined as a complete utterance that contains at least one word and does not include silence longer than 1.5 sec. We follow the definition and count up the number of turn-takings. Then, we compute the ratio for each participant.

B_3^* The ratio of the number of success interruption
B_4^* The ratio of the number of being interrupted
Same as B_1^*, leaders are also more likely to interrupt when the other participants are talking because they have to fix the way of meetings if the meetings are off the topic. In [4], success interruption for speaker S_A is defined as the situation in which S_A starts talking while another speaker S_B has already been talking, and S_B finishes his/her turn before S_A does. Rienks et al. also mentioned that the utterance by S_A had to be at least three words to eliminate backchannel noise. We utilize the same definition and count up the number of success interruption and being interrupted respectively. After that, the ratios for each participant are computed.

B_5^* The ratio of the number of large-turn-takings
B_6^* The Average time of large-turn
Turn-taking is an important clue for interaction analysis. It often happens in a multi-party meeting; namely, participants respond to each utterance and it causes too many turn-takings. Therefore, we newly define large-turn-taking, which is a duration in which a participant keeps speaking in 1.5 sec after the previous utterance by him/herself even if the other participants put some backchannels between utterances. Note that a backchannel in this experiment is an utterance that consists of less than 7 characters. It is empirically defined. We count up the number of large-turn-takings and compute the ratio for each participant. We also computed the average time of large-turns for each participant.

B. Result of the Improved Model

We compared the improved baseline with the simple baseline (B) by [3] and our method (ALL). The experimental setting was the same as the one in Section V-A. Table II shows the result of the leader identification experiment with the improved baseline. B* on the 1st column denotes a set of the new utterance features (B_1 to B_7 and B_1^* to B_6^*) in Section VI-A. Scores in bold indicate that they are higher

TABLE II
RESULTS OF LEADER IDENTIFICATION WITH THE STRONG BASELINE B*

Feature set	Leader			Non-leader		
	P	R	F1	P	R	F1
B	0.667	0.611	0.638	0.861	0.926	0.892
ALL (ours)	**0.889**	**0.833**	**0.860**	**0.935**	**1.000**	**0.967**
B*	**0.704**	**0.778**	**0.739**	**0.907**	0.889	**0.898**
B*+P	0.611	**0.722**	**0.662**	**0.907**	0.889	**0.898**
B*+F	0.426	0.611	0.502	**0.880**	0.815	0.846
B*+V	0.444	0.500	0.471	0.833	0.852	0.842
B*+P+F	0.500	0.611	0.550	**0.870**	0.852	0.861
B*+P+V	0.204	0.333	0.253	0.778	0.815	0.796
B*+F+V	0.500	**0.667**	0.571	**0.889**	0.889	0.889
B*+P+F+V	0.611	**0.722**	**0.662**	**0.898**	0.852	0.874

than the score of the baseline (B), and those with underline are the maximum values in each column.

The model with the new utterance features (B*) was actually a strong baseline. It outperformed the simple baseline model (B). This indicates that the additional utterance features (B_1^* to B_6^*) are effective for the leader identification task. On the other hand, our method (ALL), the model with all the features in Section V-B, still produced the best performance in all the criteria. This result proved the effectiveness of our method. However, the multimodal models with B* obtained lower scores than the improved baseline (B*). We need to consider the best setting of multimodal features of all. This is one important future work.

VII. CONCLUSION

In this paper, we handled the leader identification task in multi-party conversations. To generate a higher accuracy model, we proposed a leader identification model that utilizes both verbal and non-verbal features.

For the data set and verbal features, we used the setting by [3]. For non-verbal features, we extracted pose, facial, and prosodic information from video data and designed two types of pose features, four types of facial ones, and five types of prosodic ones. In our experiment, our multimodal model that contained all the features produced the best performance. We analyzed the effectiveness of our features in the model from visible results by the CART algorithm. The analysis suggested that clues from utterances and participants' facial expressions were especially important for leader identification.

Throughout the experiment, the set of utterance features played an important role, so we proposed new utterance features on the basis of the paper of Rienks et al. [4] and generated an improved baseline. The experimental result showed that our method still outperformed the improved baseline. Through the whole experiments, we concluded that it was important to consider both verbal and non-verbal features for leader identification to accomplish higher accuracy.

In this study, we conducted the experiments with a small dataset. To verify the robustness of our model, it is future work to apply our multimodal method to another large corpus. We focused on role recognition; however, our final goal is to construct an automatic behavior assessment system. To apply our method to the system is also important future work.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 20K12110.

REFERENCES

[1] Z. Ke and V. Ng, "Automated essay scoring: a survey of the state of the art," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 6300–6308.

[2] S. Okada, Y. Ohtake, Y. I. Nakano, Y. Hayashi, H.-H. Huang, Y. Takase, and K. Nitta, "Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 169–176.

[3] T. Shiota, T. Yamamura, and K. Shimada, "Analysis of facilitators' behaviors in multi-party conversations for constructing a digital facilitator system," in *Proceedings of the tenth International Conference on Collaboration Technologies*, 2018, pp. 145–158.

[4] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post, "Detection and application of influence rankings in small group meetings," in *Proceedings of the 8th International Conference on Multimodal Interfaces*, 2006, pp. 257–264.

[5] K. D. Benne and P. Sheats, "Functional roles of group members," *Journal of Social Issues*, vol. 4, no. 2, pp. 41–49, 1948.

[6] M. Zancanaro, B. Lepri, and F. Pianesi, "Automatic detection of group functional roles in face to face interactions," in *Proceedings of the 8th International Conference on Multimodal Interfaces*, 2006, pp. 28–34.

[7] H. H. Huang, Q. Zhang, S. Okada, K. Kuwabara, and T. Nishida, "Adopting functional roles for improving participants communication skill in group discussion conversation," in *Proceedings of the Group Interaction Frontiers in Technology*, 2018, pp. 1–9.

[8] T. Wilson and G. Hofer, "Using linguistic and vocal expressiveness in social role recognition," in *Proceedings of the 16th International Conference on Intelligent User Interfaces*, 2011, pp. 419–422.

[9] D. Sanchez-Cortes, O. Aran, M. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Transactions on Multimedia*, vol. 14, pp. 816–832, 2012.

[10] C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Prediction of the leadership style of an emergent leader using audio and visual nonverbal features," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 441–456, 08 2017.

[11] T. Yamamura, K. Shimada, and S. Kawahara, "The Kyutech corpus and topic segmentation using a combined method," in *Proceedings of the 12th Workshop on Asian Language Resources*, 2016, pp. 95–104.

[12] T. Yamamura, M. Hino, and K. Shimada, "Dialogue act annotation and identification in a japanese multi-party conversation corpus," in *Proceedings of the Fourth Asia Pacific Corpus Linguistics Conference*, 2018, pp. 529–536.

[13] J. D. R. Initiative, "Utterance-unit labeling manual version 2.1," 2017.

[14] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. Traum, "ISO 24617-2: A semantically-based standard for dialogue annotation." in *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 2012, pp. 430–437.

[15] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," 2018.

[16] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition*, 2018, pp. 59–66.

[17] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18 – 24.

[18] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.