MultiMediate : Multi-modal Group Behaviour Analysis for Artificial Mediation

Philipp Müller
DFKI GmbH
Saarbrücken, Germany
philipp.mueller@dfki.de

Michael Dietz*
Augsburg University
Augsburg, Germany
michael.dietz@informatik.uni-augsburg.de

Dominik Schiller*
Augsburg University
Augsburg, Germany
dominik.schiller@informatik.uni-augsburg.de

Dominike Thomas*
University of Stuttgart
Stuttgart, Germany
dominike.thomas@vis.uni-stuttgart.de

Guanhua Zhang*
University of Stuttgart
Stuttgart, Germany
guanhua.zhang@vis.uni-stuttgart.de

Patrick Gebhard
DFKI GmbH
Saarbrücken, Germany
patrick.gebhard@dfki.de

Elisabeth André
Augsburg University
Augsburg, Germany
andre@informatik.uni-augsburg.de

Andreas Bulling
University of Stuttgart
Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

ABSTRACT

Artificial mediators are promising to support human group conversations but at present their abilities are limited by insufficient progress in group behaviour analysis. The MultiMediate challenge addresses, for the first time, two fundamental group behaviour analysis tasks in well-defined conditions: *eye contact detection* and *next speaker prediction*. For training and evaluation, MultiMediate makes use of the MPIIGroupInteraction dataset consisting of 22 three- to four-person discussions as well as of an unpublished test set of six additional discussions. This paper describes the MultiMediate challenge and presents the challenge dataset including novel fine-grained speaking annotations that were collected for the purpose of MultiMediate . Furthermore, we present baseline approaches and ablation studies for both challenge tasks.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

challenge, dataset, eye contact detection, next speaker prediction

ACM Reference Format:
Philipp Müller, Michael Dietz*, Dominik Schiller*, Dominike Thomas*, Guanhua Zhang*, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2021. MultiMediate : Multi-modal Group Behaviour Analysis for Artificial Mediation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3474085.3479219

*These authors contributed equally to this work.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3479219

1 INTRODUCTION

Conversations play a central role in our lives – be it during business meetings, family gatherings, or in study groups. How people interact in conversations has a significant impact on interaction outcomes. For example, if a shy person does not speak up during a brainstorming session, valuable ideas might be overlooked, or if discussions escalate and become personal, the group may not be able to solve its tasks efficiently.

As a result, a number of artificial systems have been proposed to support human conversations and improve interaction outcomes [2, 31, 34]. One of the most ambitious ways to support humans in conversations is via artificial mediators [31] which have the advantage to resemble human interactions. Among others, artificial mediators have been studied in the context of mental health [6], education [13, 23], and collaborative teamwork [7, 36]. However, current artificial mediators still typically rely on Wizard-of-Oz paradigms to circumvent challenging sensing tasks required to adequately react to group behaviour [6, 13, 23, 35]. To realise the vision of an autonomous artificial mediator supporting group conversations, significant improvements on several fundamental group behaviour sensing and understanding tasks are required.

We introduce the MultiMediate challenge to help realise this vision by facilitating measurable progress on central group behaviour sensing and analysis tasks over several years. This year, MultiMediate focuses on *eye contact detection* and *next speaker prediction*. Both are fundamental tasks to be solved to interpret human group behaviour as well as to seamlessly interact with the group. Eye contact is linked to many important aspects of group interactions, including leadership [8, 26], interpretation of emotional facial expressions [15, p. 147] turn-taking [20, 22], and liking [22]. Similarly,

predicting who will speak next is a key human ability, enabling us to plan and start responses already before the interlocutor has finished speaking [4, 38, 39]. Apart from enabling mediators to seamlessly insert utterances, it may also allow them to proactively guide the conversation to balance participants' contributions.

In this paper, we present the first challenge for eye contact detection and next speaker prediction in group interactions that evaluates participants' approaches on an unpublished test set. We further propose baseline approaches for each challenge task and perform comprehensive evaluations. Finally, we present novel fine-grained speaking status annotations for MPIIGroupInteraction [28] that are made publicly available for future use.

2 PREVIOUS WORK

We review previous work on eye contact detection in group interactions, as well as next speaker prediction.

2.1 Eye Contact Detection

In contrast to the continuous gaze estimation task [43], eye contact detection methods use a discrete output space [27, 42]. In this way, the problem is not only simplified, but it also directly enables the computation of relevant group interaction features like the amount of gaze a speaker is receiving from interactants. Due to the difficulties of extracting meaningful information from eye regions covering only a few pixels in ambient camera recordings, past work on eye contact detection in group interactions often used head pose as a proxy to gaze direction [5, 14] or treated gaze as a latent variable [1, 30]. Recent advances in gaze estimation from standard RGB cameras [3, 43] have enabled eye contact detection in group interactions to more heavily rely on evidence extracted from the eye regions. For example, [29] presented an approach to eye contact detection using convolutional neural networks trained on head pose, utterances and horizontal eye direction. Subsequently, [41] proposed an approach based on eye gaze and head pose estimates extracted from OpenFace [3] fed into a mulitlayer perceptron. To overcome the need for annotations specific to participants' seating positions, [27] exploited the correlation between gaze and speaking behaviour in an unsupervised eye contact detection approach.

Despite these advances, eye contact detection in group interactions from RGB cameras is still far from being solved. E.g. the recent approach of [41] reaches an accuracy of 64.5% on the AMI corpus [9] and the approach of [27] reaches 63% accuracy on the MPIIGroupInteraction corpus [28]. With the MultiMediate challenge, we intend to boost research on eye contact detection by increasing attention for this challenging task and providing a fair evaluation of methods on yet unpublished data.

2.2 Next Speaker Prediction

Humans use a multitude of cues to not only plan their speaking turn and predict others' [10, 17, 20], but also to signal to others their intention of taking a turn [32]. Researchers have long attempted to understand this human skill [22] and to apply machine learning to predict turn ending and turn taking, using a wide variety of features such as mouth opening [17], head movement [16], respiration [17], blinking [11], gestures [25], word content [12] and gaze [19–21]. Most studies on conversation turn analysis focus on turn changing, ending, yielding or taking, collectively called end-of-turn estimation or turn management [17, 24], while fewer address the challenging issue of identifying *who* will speak next.

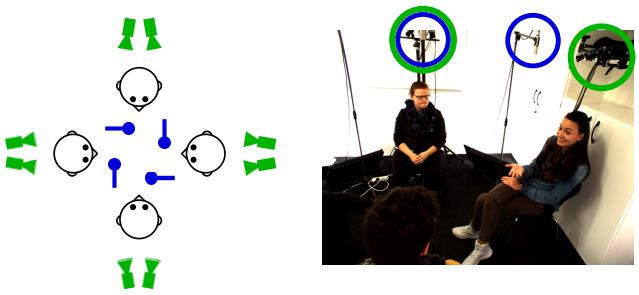

Figure 1: Illustration of the recording setup for the MPIIGroupInteraction dataset. Cameras are indicated with green, microphones with blue. Figure printed with permission from the authors of [28].

Furthermore, the majority of recent research on multi-party conversations is done on private datasets [18, 21], of which many are in Japanese [18, 21, 24]. Since turn-taking timings have been shown to vary by language [39], it is unclear how those findings transfer to other languages. The AMI corpus [9] is the only non-Japanese multi-party corpus used in a handful of next speaker prediction tasks, only one of which uses machine learning [24, 32], and mostly contains non-native speakers, which has been shown to affect conversational flow [40]. In multi-party discussions, successful uni-modal (non-verbal) models, with better than chance accuracy, predict the next speaker using head movements [16], eye-gaze [17, 18] or mouth opening [17], while the most successful recent models are multi-modal, using for example eye gaze and mouth opening [17] or eye gaze and dialogue features [24]. Similar multi-modal approaches have also been shown to be most successful in humans taking turns in human conversation, presumably because they remove ambiguities as to the intent of the next speaker [32].

While existing literature mostly used inter-pausal units [38] or dialog acts [24] as anchor points for speaker prediction, we propose to use a fixed observation time-window, as well as a fixed time point for prediction. Thereby, we remove the dependence on pause durations or turn annotations. Similar continuous models [33, 37] have so far been limited to dyads. To the best of our knowledge, MultiMediate for the first time approaches next speaker prediction in this way, on a multi-modal, multi-party dataset.

3 CHALLENGE DESCRIPTION

MultiMediate makes use of the MPIIGroupInteraction dataset [27, 28]. For next speaker prediction, we extend MPIIGroupInteraction with new fine-grained annotations of speaking status. This dataset has the distinct advantage of the availability of unpublished recordings that can be used for evaluation (see Section 3.1). We first describe MPIIGroupInteraction and subsequently discuss annotation procedures and task definitions for both challenge tasks.

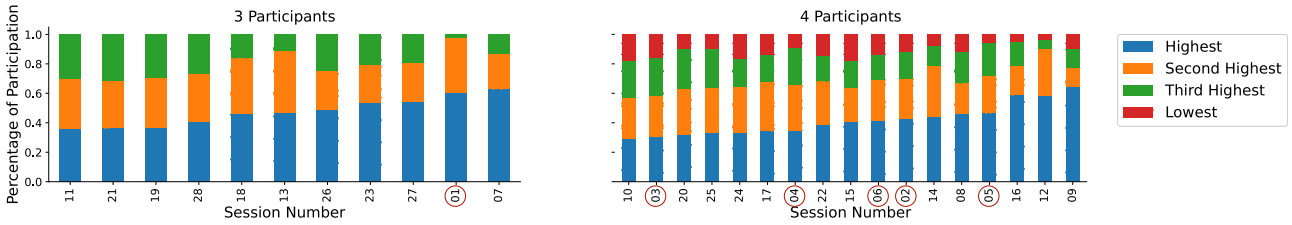

Figure 2: Distribution of the speaking time per sessions. The sessions are grouped by the number of participants. Each group is sorted in ascending order with respect to the participant that has the highest percentage of participation. Red circles indicate sessions from the test set.

3.1 Dataset

Training data. We use the already published part of MPIIGroupInteraction [28] as training data. This dataset consists of 22 conversations between three to four people, lasting 20 minutes each. For each group, the study manager chose a discussion topic that was maximally controversial among the participants. Interactants discussed this topic and were recorded by eight frame-synchronised video cameras and four microphones (see Figure 1).

Evaluation data. To test the competitors' algorithms we make use of six recordings (five four-person conversations, one three-person conversation) that were made during the creation of the MPIIGroupInteraction corpus [28] but have not yet been shared with other researchers. These recordings follow the exact same procedure as the recordings included in MPIIGroupInteraction with the minor exception that the discussion topic was not picked to be maximally controversial among the group members. Instead a discussion topic was randomly assigned to the group and group members were instructed to choose opposing opinions for themselves. This detail was changed for the final recordings of MPIIGroupInteraction in order to create more friction. However, this change does not affect the utility of these recordings for evaluating the challenge tasks.

3.2 Eye Contact Detection Task

Eye Contact Annotations. In a later study, additional eye contact annotations were collected for a total of 6,254 frames of the recordings [27]. These annotations indicate whether a participant is looking at another participant's face at a given moment in time, and if yes, who this other participant is.

Task Definition. In line with previous work [27], we define eye contact as a discrete indication of whether a participant is looking at another participants' face, and if so, who this other participant is. Video and audio recordings over a 10 second observation window are provided as temporal context for the classification decision. Eye contact has to be detected for the last frame of this window, making the task formulation also applicable to an online prediction scenario as encountered by artificial mediators. The task is modelled using five classes - one for each participants' position and an additional class for no eye contact. We use accuracy as performance metric.

3.3 Next Speaker Prediction Task

Speaking Status Annotations. We annotated all recordings with respect to the current speaker according to a strict annotation protocol. Besides speech during longer utterances, the annotators where instructed to label back-channels (e.g. "mhm" or "right") and short affirmative or dissenting statement (e.g "yes", "no") as *speaking*. Nonverbal sounds like coughing or laughing were explicitly labelled as *not speaking*, as were longer pauses during an utterance that noticeably impact the flow of speech (e.g. thinking pauses). In cases where it was still difficult to assess if a person is speaking, for example if the voice of the speaker is very quiet or sounds similar to that of another speaker, annotators were encouraged to take body language and lip movements into account. Figure 2 shows the distribution of the speaking time per session. Most sessions have one or two dominant speakers, while other participants have little speaking time. This is an ideal scenario where a virtual mediator could smoothly intervene to help balance speaking times.

Task definition. Given an observation window of 10 seconds starting at time t, participants have to predict who is speaking at time $t + 11s$, i.e. one second after the end of the observation window. We define the next speaker prediction task as a multi-label classification problem where a model should predict a binary value (speaking, not-speaking) for each participant. We evaluate performance with the unweighted average recall over all samples.

3.4 Evaluation approach

Participants are required to submit a docker image with their code to eval.ai[1] where it will be evaluated on the unpublished test set. While for each task an evaluation sample consists of a 10 second observation window of audio and video data, the sampling methods differ between tasks. For eye contact detection the available annotations consist of single frames in regular intervals [27]. We use these annotated frames as anchor points for the observation windows such that the annotated frame is the last frame of the window. For next speaker prediction we use the frames before a speaker change occurs as anchor points and subtract a random offset in the range of [0, 1000] milliseconds to determine the last frame of the observation window. In this way we ensure that the next speaker will not always start to speak exactly one second after the end of the input window, which is in line with our online prediction scenario. To balance the data set, we generated an equal number of samples with random anchor points where no speaker change occurs.

[1] https://eval.ai

Model	Featureset	Val ACC	Test ACC
Individual	Head Pose	0.51	-
	Gaze	0.48	-
	Gaze + Head Pose	**0.54**	0.52
Joint	Head Pose	0.50	-
	Gaze	0.43	-
	Gaze + Head Pose	0.53	-
Most likely class		0.33	0.26

Table 1: Accuracies for eye contact detection obtained by models trained specifically for each seating position ("Individual"), or jointly across seating positions ("Joint").

Group Features	Featureset	Val UAR	Test UAR
No	Static	0.54	-
	Dynamic	0.53	-
	Static + Dynamic	0.55	-
Yes	Static	**0.60**	-
	Dynamic	0.53	-
	Static + Dynamic	**0.60**	0.51
Most likely class		0.50	0.50

Table 2: Unweighted Average Recall (UAR) for next speaker prediction with different featurests.

4 EXPERIMENTS AND RESULTS

4.1 Eye Contact Detection

4.1.1 Method. Our baseline method makes use of features extracted via OpenFace 2.0 [3]. For a given sample, we run OpenFace on the last video frame of the input, as this is the frame for which an eye contact prediction needs to be made. We use head pose 3D (translation and rotation) as well as the 3D eye gaze direction vectors for both eyes. With these 12-dimensional feature vectors, we train separate RBF-SVMs for each seating position, resulting in four eye contact detection models. We choose γ and C parameters of the SVM by 10-fold cross-validation on the training set. For evaluation on the test set, we use both training and validation sets for training, to evaluate on the validation set we only train on the training set.

4.1.2 Results. Our baseline method achieved 0.52 accuracy on the test set. To determine the influence of features and training procedure on performance, we evaluated ablated versions of our method on the validation set (Table 1). The best result on the validation set (0.54 Accuracy) was achieved by our full baseline method. Using either only head pose or gaze features reduced accuracy to 0.51 and 0.48, respectively. We also trained a joint model across all seating positions by generating an encoding of the eye contact labels that is relative to the participant for whom we estimate the eye contact. This joint model yielded worse results for all feature sets, indicating that seating positions on the dataset are not interchangeable. All model performances are clearly above the naive baseline of always predicting the most frequent class (no eye contact) at 0.33 accuracy. At the same time, the best performance achieved in our experiments (0.54 accuracy) still leaves significant room for improvement. In comparison to previous work [27], the benefit of using gaze information is small (0.54 versus 0.51). This indicates that improvements to gaze estimation methods or the integration of gaze estimates could directly translate to improved eye contact detection.

4.2 Next Speaker Prediction

4.2.1 Method. Our method for next speaker prediction makes use of features extracted via OpenFace 2.0 [3] over the complete input video (frame by frame). We use static as well as dynamic features. Static features are head pose 3D (rotation), 3D eye gaze direction vectors for both eyes and facial action units 25 (lips part) and 26 (jaw drop) extracted from the last frame of the observation window. Dynamic features are the mean values of differences of 3D head pose (translation), AU25 and AU26 between frames, calculated over the last 4 seconds of the observation window. We train separate RBF-SVMs for each seating position to predict if the participant is the next speaker. We choose γ and C parameters of the SVM by 5-fold cross-validation on the training set. We only use the training set for training and evaluate on the validation set.

4.2.2 Results. We performed ablation experiments on the validation set to determine the influence of features (Table 2). We investigated two dimensions in our experiments: first, static features, dynamic features or their fusion; second, features from only one subject or features from all three or four subjects in the same group were used. The best result was achieved by our full baseline method using all features from all group members at 0.60 recall. However, using static features only, from all group members, also achieved a recall of 0.60. Using dynamic features only results in worse results compared to static or all features. On the testing set, our approach received a recall score of 0.51, which is lower than that on the validation set. All results on the validation set are above the trivial baseline of always predicting the most frequent class (no speaker) at 0.50 recall, but the low recall, especially on the testing set, shows that next speaker prediction remains a challenging task. The performance drop on the testing set may come from a difference in feature or label distributions between the training+validation and the testing sets. However, our results provide us with an interesting insight into group turn-taking dynamics: since group features outperform the individual ones, we can conclude that what is most relevant for next speaker prediction is not how someone behaves, but rather how they behave in comparison to the other participants.

5 CONCLUSION

We introduced MultiMediate, the first challenge addressing eye contact detection and next speaker prediction in well-defined conditions and evaluated baseline approaches for each task. In the future iterations of MultiMediate, we plan to build upon this years' achievements and add more high-level tasks that ultimately will enable machines to effectively mediate natural human interactions.

ACKNOWLEDGMENTS

P. Müller was funded by the German Ministry for Education and Research (BMBF), grant number 01IS20075. A. Bulling was funded by the European Research Council (ERC; grant agreement 801708).

REFERENCES

[1] S. O. Ba and J.-M. Odobez. 2010. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 1 (2010), 101–116. https://doi.org/10.1109/TPAMI.2010.69

[2] M. Balaam, G. Fitzpatrick, J. Good, and E. Harris. 2011. Enhancing interactional synchrony with an ambient display. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*. 867–876. https://doi.org/10.1145/1978942.1979070

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proc. IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 59–66. https://doi.org/10.1109/FG.2018.00019

[4] M. Barthel, A. S. Meyer, and S. C. Levinson. 2017. Next Speakers Plan Their Turn Early and Speak after Turn-Final "Go-Signals". *Frontiers in Psychology* 8 (2017). https://doi.org/10.3389/fpsyg.2017.00393

[5] C. Beyan, F. Capozzi, C. Becchio, and V. Murino. 2017. Prediction of the Leadership Style of an Emergent Leader Using Audio and Visual Nonverbal Features. *IEEE Transactions on Multimedia* 20, 2 (2017), 441–456. https://doi.org/10.1109/TMM.2017.2740062

[6] C. Birmingham, Z. Hu, K. Mahajan, E. Reber, and M. J. Mataric. 2020. Can I Trust You? A User Study of Robot Mediation of a Support Group. *arXiv preprint arXiv:2002.04671* (2020).

[7] D. Bohus and E. Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *Proc. International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. 1–8. https://doi.org/10.1145/1891903.1891910

[8] F. Capozzi, C. Beyan, A. Pierro, A. Koul, V. Murino, S. Livi, A. P. Bayliss, J. Ristic, and C. Becchio. 2019. Tracking the Leader: Gaze Behavior in Group Interactions. *Iscience* 16 (2019), 242–249. https://doi.org/10.1016/j.isci.2019.05.035

[9] J. Carletta, S. Ashby, S. Bourban, M. Flynn, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, and M. W. P. D. Reidsma. 2006. The ami meeting corpus: A pre-announcement. In *Proc. International Workshop on Machine Learning for Multimodal Interaction*. 28–39. https://doi.org/10.1007/11677482_3

[10] R. E. Corps, C. Gambi, and M. J. Pickering. 2018. Coordinating Utterances During Turn-Taking: The Role of Prediction, Response Preparation, and Articulation. *Discourse Processes* 55, 2 (Feb. 2018), 230–240. https://doi.org/10.1080/0163853X.2017.1330031

[11] F. Cummins. 2012. Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals. *Language and Cognitive Processes* 27, 10 (Dec. 2012), 1525–1549. https://doi.org/10.1080/01690965.2011.615220

[12] E. Ekstedt and G. Skantze. 2020. TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2981–2990. https://doi.org/10.18653/v1/2020.findings-emnlp.268

[13] O. Engwall and J. Lopes. 2020. Interaction and collaboration in robot-assisted language learning for adults. *Computer Assisted Language Learning* (2020), 1–37. https://doi.org/10.1080/09588221.2020.1799821

[14] Daniel Gatica-Perez, L. McCowan, Dong Zhang, and Samy Bengio. 2005. Detecting group interest-level in meetings. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. IEEE, I–489. https://doi.org/10.1109/ICASSP.2005.1415157

[15] Ursula Hess and Agneta Fischer. 2013. Emotional Mimicry as Social Regulation. *Personality and Social Psychology Review* 17, 2 (2013), 142–157. https://doi.org/10.1177/1088868312472607

[16] R. Ishii, S. Kumano, and K. Otsuka. 2015. Predicting next speaker based on head movement in multi-party meetings. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. 2319–2323. https://doi.org/10.1109/ICASSP.2015.7178385

[17] R. Ishii, K. Otsuka, S. Kumano, R. Higashinaka, and J. Tomita. 2019. Prediction of Who Will Be Next Speaker and When Using Mouth-Opening Pattern in Multi-Party Conversation. *Multimodal Technologies and Interaction* 3, 4 (Dec. 2019), 70. https://doi.org/10.3390/mti3040070

[18] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Masafumi Matsuda, and Junji Yamato. 2013. Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In *Proc. ACM International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 79–86. https://doi.org/10.1145/2522848.2522856

[19] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato. 2016. Prediction of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings. *ACM Transactions on Interactive Intelligent Systems* 6, 1 (May 2016), 1–31. https://doi.org/10.1145/2757284

[20] K. Jokinen, H. Furukawa, M. Nishida, and S. Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems* 3, 2 (Aug. 2013), 12:1–12:30. https://doi.org/10.1145/2499474.2499481

[21] T. Kawahara, T. Iwatate, and K. Takanashi. 2012. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. *Proc. Annual Conference of the International Speech Communication Association* 1 (2012), 726–729.

[22] A. Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26 (Jan. 1967), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

[23] J. Lopes, O. Engwall, and G. Skantze. 2017. A first visit to the robot language café. In *Proc. ISCA workshop on Speech and Language Technology in Education*. https://doi.org/10.1007/s12369-020-00635-y

[24] U. Malik, J. Saunier, K. Funakoshi, and A. Pauchet. 2020. Who Speaks Next? Turn Change and Next Speaker Prediction in Multimodal Multiparty Interaction. In *Proc. IEEE International Conference on Tools with Artificial Intelligence*. 349–354. https://doi.org/10.1109/ICTAI50040.2020.00062

[25] L. Mondada. 2007. Multimodal resources for turn-taking: pointing and the emergence of possible next speakers. *Discourse Studies* 9 (2007), 194–225. Issue 2. https://doi.org/10.1177/1461445607075346

[26] P. Müller and A. Bulling. 2019. Emergent Leadership Detection Across Datasets. In *Proc. International Conference on Multimodal Interaction*. 274–278. https://doi.org/10.1145/3340555.3353721

[27] P. Müller, M. X. Huang, X. Zhang, and A. Bulling. 2018. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 1–10. https://doi.org/10.1145/3204493.3204549

[28] P. Müller, M. X. Huang, and A. Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *Proc. International Conference on Intelligent User Interfaces*. Association for Computing Machinery, Tokyo, Japan, 153–164. https://doi.org/10.1145/3172944.3172969

[29] K. Otsuka, K. Kasuga, and M. Köhler. 2018. Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In *Proc. ACM International Conference on Multimodal Interaction*. 191–199. https://doi.org/10.1145/3242969.3242973

[30] K. Otsuka, Y. Takemae, and J. Yamato. 2005. A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of the 7th international conference on Multimodal interfaces*. 191–198. https://doi.org/10.1145/1088463.1088497

[31] S. Park and Y. Lim. 2020. Investigating User Expectations on the Roles of Family-shared AI Speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3313831.3376450

[32] V. Petukhova and H. Bunt. 2009. 'Who's next? Speaker-selection mechanisms in multiparty dialogue'. In *DiaHolmia*. Stockholm, Sweden, 19–26.

[33] M. Roddy, G. Skantze, and N. Harte. 2018. Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs. In *Proc. ACM International Conference on Multimodal Interaction*. ACM, 186–190. https://doi.org/10.1145/3242969.3242997

[34] G. Schiavo, A. Cappelletti, E. Mencarini, O. Stock, and M. Zancanaro. 2014. Overt or subtle? Supporting group conversations with automatically targeted directives. In *Proc. International Conference on Intelligent User Interfaces*. 225–234. https://doi.org/10.1145/2557500.2557507

[35] S. Sebo, B. Stoll, B. Scassellati, and M. F. Jung. 2020. Robots in groups and teams: a literature review. *Proc. ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–36. https://doi.org/10.1145/3415247

[36] E. Short and M. J. Mataric. 2017. Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions. In *Proc. IEEE International Symposium on Robot and Human Interactive Communication*. 385–390. https://doi.org/10.1109/ROMAN.2017.8172331

[37] G. Skantze. 2017. Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks. In *Proc. Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Saarbrücken, Germany, 220–230. https://doi.org/10.18653/v1/W17-5527

[38] G. Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67 (May 2021), 101178. https://doi.org/10.1016/j.csl.2020.101178

[39] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon, and S. C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106, 26 (June 2009), 10587–10592. https://doi.org/10.1073/pnas.0903616106

[40] I. Umata, K. Ijuin, T. Kato, and S. Yamamoto. 2019. Floor Apportionment Function of Speaker's Gaze in Grounding Acts. In *Adjunct Proc. International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3351529.3360660

[41] L. Zhang, M. Morgan, I. Bhattacharya, M. Foley, J. Braasch, C. Riedl, B. Foucault Welles, and R. J. Radke. 2019. Improved visual focus of attention estimation and prosodic features for analyzing group interactions. In *Proc. ICMI*. 385–394. https://doi.org/10.1145/3340555.3353761

[42] X. Zhang, Y. Sugano, and A. Bulling. 2017. Everyday eye contact detection using unsupervised gaze target discovery. In *Proc. ACM Symposium on User Interface Software and Technology*. 193–203. https://doi.org/10.1145/3126594.3126614

[43] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. 2017. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 1 (2017), 162–175. https://doi.org/10.1109/TPAMI.2017.2778103