Published: 10 October 2022

Citation in BibTeX format

MM '22: The 30th ACM International Conference on Multimedia
October 10 - 14, 2022
Lisboa, Portugal

Conference Sponsors:
SIGMM

RESEARCH-ARTICLE

MultiMediate'22: Backchannel Detection and Agreement Estimation in Group Interactions

PHILIPP MÜLLER, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Rheinland-Pfalz, Germany

MICHAEL DIETZ, University of Augsburg, Augsburg, Bayern, Germany

DOMINIK SCHILLER, University of Augsburg, Augsburg, Bayern, Germany

DOMINIKE THOMAS, University of Stuttgart, Stuttgart, Baden-Wurttemberg, Germany

HALI LINDSAY, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Rheinland-Pfalz, Germany

PATRICK GEBHARD, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Rheinland-Pfalz, Germany

View all

Open Access Support provided by:

University of Augsburg

German Research Center for Artificial Intelligence (DFKI)

University of Stuttgart

MM '22: Proceedings of the 30th ACM International Conference on Multimedia (October 2022)
https://doi.org/10.1145/3503161.3551589
ISBN: 9781450392037

MultiMediate '22: Backchannel Detection and Agreement Estimation in Group Interactions

Philipp Müller
German Research Center for Artificial Intelligence
Saarbrücken, Germany
philipp.mueller@dfki.de

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

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

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

Hali Lindsay
German Research Center for Artificial Intelligence
Saarbrücken, Germany
hali.lindsay@dfki.de

Patrick Gebhard
German Research Center for Artificial Intelligence
Saarbrücken, Germany
patrick.gebhard@dfki.de

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

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

ABSTRACT

Backchannels, i.e. short interjections of the listener, serve important meta-conversational purposes like signifying attention or indicating agreement. Despite their key role, automatic analysis of backchannels in group interactions has been largely neglected so far. The MultiMediate challenge addresses, for the first time, the tasks of *backchannel detection* and *agreement estimation from backchannels* in group conversations. This paper describes the MultiMediate challenge and presents a novel set of annotations consisting of 7234 backchannel instances for the MPIIGroupInteraction dataset. Each backchannel was additionally annotated with the extent by which it expresses agreement towards the current speaker. In addition to a an analysis of the collected annotations, we present baseline results for both challenge tasks.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**.

KEYWORDS

challenge, dataset, backchannel detection, agreement estimation

ACM Reference Format:
Philipp Müller, Michael Dietz*, Dominik Schiller*, Dominike Thomas*, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2022. MultiMediate '22: Backchannel Detection and Agreement Estimation in Group Interactions. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3503161.3551589

*These authors contributed equally to this work.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3551589

1 INTRODUCTION

Backchannels, i.e. short interjections made by listeners, are at the core of the bilateral nature of dialogue [24]. That is, dialogue in which a listener's responses affect the speaker's speech acts. They may consist of a verbal phrase like "oh yes", a sound (e.g. "hm"), or bodily gestures, such as head nods or hand movements, or combinations thereof. Backchanneling serves the important functions of confirming listeners' attention and comprehension [33], as well as regulating turn-taking [60]. In addition, it is used to communicate agreement or disagreement with the current speaker and is therefore an important indicator of participants' opinions and intra-group relations [12]. The inability to appropriately perform backchanneling can have severe impact on dialogue, e.g. by distracting the speaker [10, 51].

As a result, automatic analysis of backchanneling behaviour is crucial and has significant potential for artificial systems designed to passively monitor or actively mediate human dialogue [6, 52, 53, 61]. For example, artificial mediators [14, 17, 25] could analyse the frequency of backchannels to gauge participants' engagement and encourage those who are disengaged from the discussion. Agreement or disagreement expressed in backchannels could also help artificial mediators to better understand the opinions of participants. This may enable mediators to give a voice to participants who find it difficult to express diverging opinions. Despite this potential, automatic analysis of backchannels in group interactions remains largely unexplored due to the lack of suitable datasets.

With MultiMediate '22, we present the first challenge for automatic backchannel analysis in group interactions. To this end, we introduce the first publicly available dataset of backchanneling behaviour in group discussions. We fully annotated the MPIIGroupInteraction dataset [49] with 7234 backchannel instances. Furthermore, each backchannel instance was rated with the extent to which it expresses agreement with the current speaker. We present analyses of our novel annotations as well as evaluations of baseline approaches for MultiMediate '22. All collected annotations, baseline implementations, and raw feature representations are made publicly available for further use.[1]

2 PREVIOUS WORK

2.1 Backchannel and Agreement Datasets

While a number of different datasets were recorded for the purpose of backchannel analysis [10, 22, 44, 55], they are rarely public and not sufficient to ensure progress on automatic backchannel analysis in group interactions. In Table 1 we list recent publicly available datasets that were used in BC research, the majority of which consists of spoken dialogues. Note that while these datasets are publicly available, the backchannel annotations collected on them are often not (column "Pub." in Table 1). Furthermore, all existing datasets (except the Canal9 Corpus [54, 67] for which no publicly available backchannel annotations exist) consist of dyadic interactions and are thus not fitting for group behaviour analysis.

In terms of annotations of backchannels, some studies use semi-automatic labeling of potential events [9, 18], or query a dataset for specific BC keywords [50]. Backchannel annotations are most often based on a set list of backchannel events ("is there a "uh-huh" here?") [18, 27, 45, 50, 63], and only rarely on the holistic perception of a backchannel ("is there a backchannel here?") [5, 11, 35].

While a relationship between backchannels and agreement has been discussed [60], few studies have systematically investigated it. [55] included agreement in their annotations of backchannels, but did not analyse it; [30, 36] distinguish backchannels from agreement signals. Some studies of agreement include backchannels [19, 20], but to the best of our knowledge, no study of backchanneling investigated agreement transmitted *through* the backchannels themselves.

In our work, we present the first publicly available annotations of backchannel occurances and agreements expressed via backchannels in group interactions. Our full dataset is manually annotated and with over 33 hours of annotated human behaviour across training, validation and test sets, its size is equal to the largest dyadic BC dataset currently available (see Table 1).

2.2 Computational Models

The prediction (i.e. anticipation) of backchannels is a highly active area of research in social signal processing. One major motivation is the goal to generate natural backchanneling behaviour in artificial agents. There is a variety of traditional machine learning methods used for predicting backchannels [27, 29, 31, 40, 44, 64]. Recently, LSTM networks became the most frequent choice [1, 2, 34, 58], along with residual networks [31]. Multitask learning also seems to be a particularly successful approach [34, 37, 39]. The most common features used in BC prediction are prosodic [2, 34, 57, 58] while some research also makes use of linguistic features such as word embeddings [1, 2, 57, 58] or syntactic (part of speech tags), semantic (concreteness, valence) and discourse features [16, 18, 40]. Research shows that adding lexical features to accoustic ones improves results [50, 57]. [33] offers a review of the most important backchannel-inviting cues which may also be used as features.

Name	Pub.	Part.	Size	Lang.
Cheese-Paco [16, 18]	✗	2	2h	FR
Vyaktitv [38]	✗	2	14h	HI
Spontal [35]	✗	2	0h40	SV
P2PSTORY [62]	✓	2	2h30	EN
Canal9 [54]	✗	5	–	FR
Cup of CoFee [55]	✓	2	33h42	FR
IFADV [65]	✗	2	9h30	NL
Spoken Language [4, 5]	✗	2	–	SV
NOXI [22]	✗	2	25h18	7 Lgs
MPIIGroupInteraction [49]	✓	3-4	33h40	DE

Table 1: Publicly available audio-visual human-human interaction datasets used in BC research. *Pub.* indicates whether BC annotations are publicly available; *Part.* the number of participants per interaction. *Size* is the duration of individual human behaviour annotated with BCs (where reported). E.g. for a dyadic dataset it is twice the length of annotated interactions. *Lang.* indicates the language of the dataset.

In contrast to backchannel prediction, we define the task of backchannel detection as categorizing observed behaviour as to whether a backchannel was shown or not. While to the best of our knowledge, this task was not studied in isolation in previous work, backchannel detection can appear as a part of the multi-class problem of dialogue act classification (DAC) [3, 59]. DAC is primarily addressed by text analysis [32, 42, 56] and only few works incorporated multi-modal information like emotional expressions [21, 59]. Importantly, no previous work addressed the task of backchannel detection in group interactions from multi-modal behaviour.

With MultiMediate '22 we aim to attract researchers to two challenging problems centered around backchannels in group discussions: backchannel detection and estimation of the amount of agreement expressed in a backchannel. In line with human studies that underline the importance of non-verbal backchannel cues [5, 41], we provide annotations of backchannel behaviour that can take place both via speech and in the visual domain. We contribute to improved comparability between approaches by using an unpublished test set for evaluation.

3 CHALLENGE DESCRIPTION

As in MultiMediate '21 [47], our challenge is based on the MPIIGroupInteraction dataset [48, 49]. This dataset has been used for diverse tasks, including low rapport detection [49], emergent leadership detection [46], eye contact detection [28, 48], next speaker prediction [15], and body language detection [7]. For MultiMediate '22 we collected novel backchannel annotations on the whole dataset. Test samples (excluding ground truth) are released to participants before the challenge deadline. Participants in turn submit

[1] https://multimediate-challenge.org

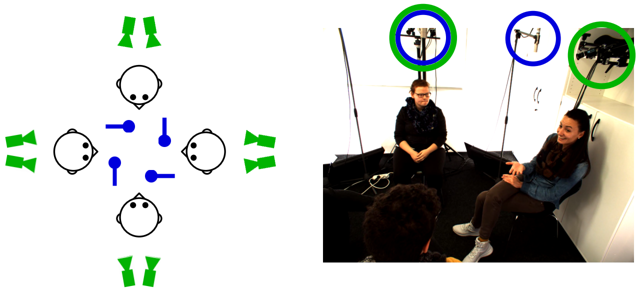

Figure 1: The recording setup for MPIIGroupInteraction. Printed with permission from the authors of [49].

their predictions for evaluation by the challenge organisers. We first describe MPIIGroupInteraction and subsequently discuss annotation procedures and task definitions for backchannel detection and agreement estimation.

3.1 Dataset

Training data. We use the publicly available recordings of MPIIGroupInteraction [49] as training data. The dataset comprises 22 conversations between three to four people on a maximally controversial topic, lasting 20 minutes each. While discussing the topic, interactants were recorded with four microphones and eight frame-synchronised video cameras (see Figure 1).

Evaluation data. We follow MultiMediate '21 [47] and use six yet unpublished discussions that were recorded during the creation of MPIIGroupInteraction [49] for testing. These six discussions followed the same procedure as the training discussions with the only exception that the topic was not chosen to be maximally controversial. Instead, a topic was randomly selected for each group and participants were asked to take on opposing views for themselves.

3.2 Backchannel Detection Task

Backchannel Annotations. In a first step annotators where asked to label the occurrences of backchanneling behaviour with respect to different modalities: *auditive*, and *visual*. Visual annotations take only the aspects observable in the video into account (e.g. nodding or head shaking). Auditive backchanneling behaviour, in turn, relies only on cues from the audio signal, encompassing verbal (e.g. "yes", "oh really?") and paraverbal behaviour (e.g. "uhm hum", "aha"). Each of three annotators labeled a specific portion of the dataset with respect to backchannels for each modality. Figure 2 (left) shows the average of annotated backchannel events annotated per participant in each modality. On average, we observed 28 annotations per participant for the auditive modality, and 60 for the visual modality. Figure 2 (middle) shows the average duration of annotated backchannel events for each modality. Visual backchannel events are on average longer (1.8 seconds) compared to auditive backchannel events (0.56 seconds). The modality specific labels were subsequently joined using the logical *OR* operator. Hence, overlapping backchannels in different modalities (e.g. a person nods while simultaneously saying "yes") are merged to a new, modality independent *backchannel* label. To provide negative examples we also calculated an equal amount of non-backchannel samples per session. Overall this results in 14468 labels partitioned into the following splits: 6716 Train, 2854 Val, 4898 Test.

Task definition. Given an observation window of 10 seconds participants have to detect if there is a backchannel present in the sample. Every sample has been created such that the end of an annotated backchannel has been used as the end the sample or that a sample consists of 10 seconds with no annotated backchannels. The performance metric for the task is accuracy.

3.3 Backchannel Agreement Estimation Task

Agreement Annotations. In the second step, all backchannel instances that where annotated for the *backchannel detection test* were labeled with respect to their level of expressed agreement on a scale from -1 (total disagreement) to 1 (total agreement) using a step size of 0.1. In case annotators found that the instance was erroneously labeled as a backchannel, they indicated this fact with an extra label. When all annotators agreed that an instance was wrongly labeled as a backchannel the respective sample has been removed from the dataset during the sampling process. This annotation step has been performed by each annotator for all backchannel labels (see section 3.2) The groundtruth labels were then created by averaging all three annotations per sample. Overall this results in 7234 labels partitioned into the following splits: 3358 Train, 1427 Val, 2449 Test. To quantify the reliability of agreement annotations, we compute the average Spearman ρ when comparing a left-out annotator to the average of the two remaining ones is. We observe a ρ of 0.62, indicating substantial agreement. Figure 2 (right) shows the distribution of aggregated agreement annotations. The distribution is centered in the positive range between 0 and 0.5. The detection of disagreement represents a special challenge due to the small number of samples below 0.

Task definition. For the backchannel agreement estimation tasks participants have to predict the average expressed agreement per sample on scale from -1 to 1. The performance metric for the task is mean squared error.

4 EXPERIMENTS AND RESULTS

4.1 Features

We extracted the same set of features for backchannel detection and agreement estimation from backchannels. Features were extracted from the last second of the 10 second input window and aggregated by computing the mean or the mean of absolute differences ("mean delta") of adjacent frames over this second.

4.1.1 Head Features. We extracted features from participants' head and face using OpenFace 2.0 [8]. These include mean and mean delta of AU intensity estimates, mean delta of head orientation and translation, as well as mean delta of gaze angles for both eyes. In total, we extracted 46 features based on OpenFace 2.0.

4.1.2 Pose Features. We extract body pose estimates using OpenPose [23] and employ a set of angular features that proved successful for group interaction analysis [13]. These features consist of angles between body parts, e.g. the angle defined by the line between left

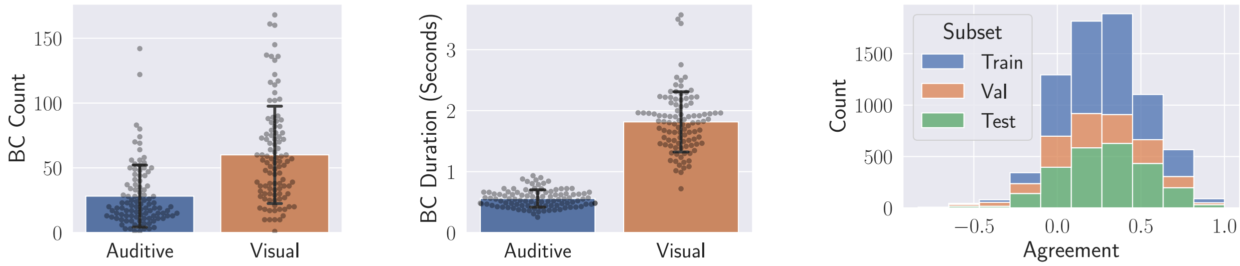

Figure 2: Left: Number of auditive and visual backchannel annotations per participant. Each point represents one participant, bars and whiskers represent the mean and standard deviation across participants. Middle: Average duration of auditive and visual backchannels. Right: Distribution of agreement scores across MultiMediate '22 train, val, and test samples.

shoulder and left elbow and the line between left shoulder and right shoulder. We limit ourselves to the "Upper Body" and "Head" features described in [13], as lower body pose estimates tend to be unreliable on the MPIIGroupInteraction dataset. We compute the mean delta of these features, resulting in 8 body pose features.

4.1.3 Voice Features. We extracted the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [26] on the last second of the input window. This set consists of 88 acoustic parameters that are commonly applied to tasks like depression, mood, and emotion recognition [66], or Alzheimer's Dementia recognition [43].

4.2 Prediction Approach

For backchannel detection, we trained a binary Support Vector Classifier (SVC) with rbf kernel. For the agreement estimation task, we trained a Support Vector Regressor (SVR) with rbf kernel. We employed 10-fold cross-validation on the training set to choose γ and C parameters of the SVC/SVR. For test set evaluations, we trained on training and validation sets; for evaluations on the validation set we only trained on the MultiMediate '22 training set.

4.3 Results

We present evaluation results for different combinations of feature sets in Table 2. For fairness reasons, we only evaluate two feature sets for each task on the training set: the best performing feature set on the validation set, as well as all included features.

4.3.1 Backchannel Detection. The best feature set on the validation set consisted of a combination of head and pose features, reaching 0.639 accuracy on validation and 0.596 accuracy on the test set. This clearly outperformed the trivial baseline of a random predictor at 0.5 accuracy and was marginally better than all featuresets combined. Notably, our experiments on the validation set revealed that each individual feature set achieved above-random performance on backchannel detection, even though OpenFace 2.0 based head features were clearly leading. An ablation of the head feature set revealed that head pose alone (i.e. mean delta of translation and rotation of the head) reaches 0.636 accuracy. This is likely due to the strong association between backchanneling and nodding.

4.3.2 Agreement Estimation. For agreement estimation, head pose features performed best on the validation set (0.075 MSE) and reached 0.061 MSE on the test set. In both cases this improves above the trivial baseline of using the mean on the training set as a predictor. However, featuresets that do not include head pose ore gaze features fail to outperfrom the trivial baseline. This is in contrast to the backchannel detection task and indicates the difficulty of backchannel agreement estimation.

Features	Detection Val ACC ↑	Detection Test ACC ↑	Agreement Val MSE ↓	Agreement Test MSE ↓
Head	0.621	-	0.079	-
AUs only	0.591	-	0.085	-
H. Pose only	0.636	-	**0.075**	**0.061**
Gaze only	0.622	-	0.078	-
Pose	0.531	-	0.086	-
Voice	0.567	-	0.085	-
Head + Pose	**0.639**	**0.596**	0.079	-
All Features	0.636	0.592	0.079	0.064
Trivial Basel.	0.500	0.500	0.085	0.066

Table 2: Validation and test results for backchannel detection and agreement estimation from backchannels.

5 CONCLUSION

We introduced MultiMediate '22, the first challenge addressing backchannel detection and agreement estimation from backchannels in well-defined conditions and evaluated baseline approaches for each task. In addition we introduced a novel publicly available dataset of backchannel annotations in group interactions that is a valuable resource for research on backchannel detection and agreement estimation, even beyond the MultiMediate challenge.

ACKNOWLEDGMENTS

P. Müller and H. Lindsay were 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). The researchers from Augsburg University were partially funded by the German Ministry for Education and Research (BMBF), grant numbers 16SV8688 and 16DHB2215.

REFERENCES

[1] Amalia Istiqlali Adiba, Takeshi Homma, Dario Bertero, Takashi Sumiyoshi, and Kenji Nagamatsu. 2021. Delay Mitigation for Backchannel Prediction in Spoken Dialog System. In *Conversational Dialogue Systems for the Next Decade*. Vol. 704. Springer, 129–143. https://doi.org/10.1007/978-981-15-8395-7_10

[2] Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori Miyoshi. 2021. Towards Immediate Backchannel Generation Using Attention-Based Early Prediction Model. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 7408–7412. https://doi.org/10.1109/ICASSP39728.2021.9414193

[3] Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1997. Dialogue acts in VERBMOBIL-2. (1997).

[4] Jens Allwood. 1999. The Swedish Spoken Language Corpus at Göteborg University. In *Proc. Fonetik 99: The Swedish Phonetics Conf.* Göteborg University, Sweden, 1–5.

[5] Jens Allwood and Loredana Cerrato. 2003. A study of gestural feedback expressions. In *Proc. of the Nordic Symposium on Multimodal Communication*. 23–24.

[6] Madeline Balaam, Geraldine Fitzpatrick, Judith Good, and Eric Harris. 2011. Enhancing interactional synchrony with an ambient display. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. 867–876. https://doi.org/10.1145/1978942.1979070

[7] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and François Brémond. 2022. Bodily Behaviors in Social Interaction: Novel Annotations and State-of-the-Art Evaluation. In *Proc. of the ACM International Conference on Multimedia*. https://doi.org/10.1145/3503161.3548363

[8] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proc. of the IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 59–66. https://doi.org/10.1109/FG.2018.00019

[9] Tobias Baur, Alexander Heimerl, Florian Lingenfelser, Johannes Wagner, Michel F. Valstar, Björn W. Schuller, and Elisabeth André. 2020. eXplainable Cooperative Machine Learning with NOVA. *Künstliche Intell.* 34, 2 (2020), 143–164. https://doi.org/10.1007/s13218-020-00632-3

[10] Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology* 79, 6 (2000), 941–952. https://doi.org/10.1037/0022-3514.79.6.941

[11] Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2002. Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication* 52, 3 (2002). https://doi.org/10.1111/j.1460-2466.2002.tb02562.x

[12] Elisabetta Bevacqua, Sathish Pammi, Sylwia Julia Hyniewska, Marc Schröder, and Catherine Pelachaud. 2010. Multimodal Backchannels for Embodied Conversational Agents. In *Intelligent Virtual Agents*. Vol. 6356. Springer, 194–200. https://doi.org/10.1007/978-3-642-15892-6_21

[13] Cigdem Beyan, Vasiliki-Maria Katsageorgiou, and Vittorio Murino. 2017. Moving as a leader: Detecting emergent leadership in small groups using body pose. In *Proc. of ACM International Conference on Multimedia*. 1425–1433. https://doi.org/10.1145/3123266.3123404

[14] Chris Birmingham, Zijian Hu, Kartik Mahajan, Eli Reber, and Maja J. Mataric. 2020. Can I Trust You? A User Study of Robot Mediation of a Support Group. *arXiv preprint arXiv:2002.04671* (2020).

[15] Chris Birmingham, Kalin Stefanov, and Maja J Mataric. 2021. Group-Level Focus of Visual Attention for Improved Next Speaker Prediction. In *Proc. of the 29th ACM International Conference on Multimedia*. 4838–4842. https://doi.org/10.1145/3474085.3479213

[16] Philippe Blache, Massina Abderrahmane, Stéphane Rauzy, and Roxane Bertrand. 2020. An integrated model for predicting backchannel feedbacks. In *Proc. of the ACM International Conference on Intelligent Virtual Agents*. 1–3. https://doi.org/10.1145/3383652.3423948

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

[18] Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. 2021. A Multimodal Model for Predicting Conversational Feedbacks. In *Text, Speech, and Dialogue (Lecture Notes in Computer Science)*. Springer International Publishing, 537–549. https://doi.org/10.1007/978-3-030-83527-9_46

[19] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. 2009. Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *Proc. of the International Conference on Affective Computing and Intelligent Interaction and Workshops*. 1–9. https://doi.org/10.1109/ACII.2009.5349477

[20] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. 2013. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing* 31, 2 (Feb. 2013), 203–221. https://doi.org/10.1016/j.imavis.2012.07.003

[21] Kristy Boyer, Joseph F Grafsgaard, Eun Young Ha, Robert Phillips, and James Lester. 2011. An affect-enriched dialogue act classification model for task-oriented dialogue. In *Proc. of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 1190–1199.

[22] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres, Catherine Pelachaud, Elisabeth André, and Michel F. Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proc. of the ACM International Conference on Multimodal Interaction*. ACM, 350–359. https://doi.org/10.1145/3136755.3136780

[23] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299. https://doi.org/10.1109/CVPR.2017.143

[24] Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50, 1 (Jan. 2004), 62–81. https://doi.org/10.1016/j.jml.2003.08.004

[25] Olov Engwall and José 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

[26] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2015), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

[27] Gaëlle Ferré and Suzanne Renaudier. 2017. Unimodal and Bimodal Backchannels in Conversational English. In *Proc. of Workshop on the Semantics and Pragmatics of Dialogue*. 27–37. https://doi.org/10.21437/SemDial.2017-3

[28] Eugene Yujun Fu and Michael W Ngai. 2021. Using Motion Histories for Eye Contact Detection in Multiperson Group Conversations. In *Proc. of the ACM International Conference on Multimedia*. 4873–4877. https://doi.org/10.1145/3474085.3479230

[29] Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. 2005. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *Proc. Interspeech*. ISCA, 889–892. https://doi.org/10.21437/Interspeech.2005-400

[30] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*. 669–676. https://doi.org/10.3115/1218955.1219040

[31] Mononito Goswami, Minkush Manuja, and Maitree Leekha. 2020. Towards Social & Engaging Peer Learning: Predicting Backchanneling and Disengagement in Children. *arXiv:2007.11346* (2020).

[32] Sergio Grau, Emilio Sanchis, Maria Jose Castro, and David Vilar. 2004. Dialogue act classification using a Bayesian approach. In *Proc. of the Conference Speech and Computer*.

[33] Agustín Gravano and Julia Hirschberg. 2009. Backchannel-inviting cues in task-oriented dialogue. In *Proc. Interspeech*. ISCA, 1019–1022. https://doi.org/10.21437/Interspeech.2009-301

[34] Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Prediction of Turn-taking Using Multitask Learning with Prediction of Backchannels and Fillers. In *Proc. Interspeech*. ISCA, 991–995. https://doi.org/10.21437/Interspeech.2018-1442

[35] Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. 2013. Backchannel relevance spaces. In *Proc. of Nordic Prosody XI*. 137–146.

[36] Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection Of Agreement vs. Disagreement In Meetings: Training With Unlabeled Data. In *Proc. of the HLT-NAACL (Companion)*. 34–36. https://doi.org/10.3115/1073483.1073495

[37] Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2021. Multimodal and Multitask Approach to Listener's Backchannel Prediction: Can Prediction of Turn-changing and Turn-management Willingness Improve Backchannel Modeling?. In *Proc. of the ACM International Conference on Intelligent Virtual Agents*. 131–138. https://doi.org/10.1145/3472306.3478360

[38] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring Semi-Supervised Learning for Predicting Listener Backchannels. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. Number 395. 1–12. https://doi.org/10.1145/3411764.3445449

[39] Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. 2021. BPM_MT: Enhanced Backchannel Prediction Model using Multi-Task Learning. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*. 3447–3452. https://doi.org/10.18653/v1/2021.emnlp-main.277

[40] Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel Ward. 2016. Prediction and Generation of Backchannel Form for Attentive Listening Systems. In *Proc. Interspeech*. ISCA, 2890–2894. https://doi.org/10.21437/Interspeech.2016-118

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

[42] Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proc. of the International Conference on Computational Linguistics: Technical Papers*. 2012–2021.

[43] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833* (2020).

[44] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2008. Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*. Springer, 176–190. https://doi.org/10.1007/978-3-540-85483-8_18

[45] Markus Mueller, David Leuschner, Lars Briem, Maria Schmidt, Kevin Kilgour, Sebastian Stueker, and Alex Waibel. 2015. Using Neural Networks for Data-Driven Backchannel Prediction: A Survey on Input Features and Training Techniques. In *Human-Computer Interaction: Interaction Technologies (Lecture Notes in Computer Science)*. Springer International Publishing, 329–340. https://doi.org/10.1007/978-3-319-20916-6_31

[46] Philipp Müller and Andreas Bulling. 2019. Emergent Leadership Detection Across Datasets. In *Proc. of the International Conference on Multimodal Interaction*. 274–278. https://doi.org/10.1145/3340555.3353721

[47] 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 *Proc. of the ACM International Conference on Multimedia*. 4878–4882. https://doi.org/10.1145/3474085.3479219

[48] Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. 2018. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proc. of the ACM Symposium on Eye Tracking Research & Applications*. 1–10. https://doi.org/10.1145/3204493.3204549

[49] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *Proc. of the ACM International Conference on Intelligent User Interfaces*. 153–164. https://doi.org/10.1145/3172944.3172969

[50] Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. OH, JEEZ! or UH-HUH? A Listener-Aware Backchannel Predictor on ASR Transcriptions. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 8064–8068. https://doi.org/10.1109/ICASSP40776.2020.9054223

[51] Hae Won Park, Mirko Gelsomini, Jin Joo Lee, Tonghui Zhu, and Cynthia Breazeal. 2017. Backchannel opportunity prediction for social robot listeners. In *Proc. of the IEEE International Conference on Robotics and Automation*. 2308–2314. https://doi.org/10.1109/ICRA.2017.7989266

[52] Sunjeong Park and Youn-kyung Lim. 2020. Investigating User Expectations on the Roles of Family-shared AI Speakers. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3313831.3376450

[53] Anna Penzkofer, Philipp Müller, Felix Bühler, Sven Mayer, and Andreas Bulling. 2021. ConAn: A Usable Tool for Multimodal Conversation Analysis. In *Proc. of the International Conference on Multimodal Interaction*. 341–351. https://doi.org/10.1145/3462244.3479886

[54] Isabella Poggi, Francesca D'Errico, and Laura Vincze. 2010. Types of Nods. The Polysemy of a Social Signal. In *Proc. of the International Conference on Language Resources and Evaluation*.

[55] Laurent Prévot, Jan Gorisch, and Roxane Bertrand. 2016. A CUP of CoFee: A large Collection of feedback Utterances Provided with communicative function annotations. In *Proc. of the International Conference on Language Resources and Evaluation*. 3180–3185.

[56] Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv:1904.02594* (2019).

[57] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing Backchannel Prediction Using Word Embeddings. In *Proc. Interspeech*. ISCA, 879–883. https://doi.org/10.21437/Interspeech.2017-1606

[58] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. Yeah, Right, Uh-Huh: A Deep Learning Backchannel Predictor. In *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*. Springer International Publishing, 247–258. https://doi.org/10.1007/978-3-319-92108-2_25

[59] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards emotion-aided multi-modal dialogue act classification. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*. 4361–4372.

[60] Emanuel A. Schegloff. 1982. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In *Analyzing Discourse: Text and Talk*. Georgetown University Press, 71–93.

[61] Gianluca Schiavo, Alessandro Cappelletti, Eleonora Mencarini, Oliviero Stock, and Massimo Zancanaro. 2014. Overt or subtle? Supporting group conversations with automatically targeted directives. In *Proc. of the International Conference on Intelligent User Interfaces*. 225–234. https://doi.org/10.1145/2557500.2557507

[62] Nikhita Singh, Jin Joo Lee, Ishaan Grover, and Cynthia Breazeal. 2018. P2PSTORY: Dataset of Children as Storytellers and Listeners in Peer-to-Peer Interactions. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. 1–11. https://doi.org/10.1145/3173574.3174008

[63] Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. 2014. Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication* 65 (2014), 50–66. https://doi.org/10.1016/j.specom.2014.05.005

[64] Allison Terrell and Bilge Mutlu. 2012. A Regression-based Approach to Modeling Addressee Backchannels. In *Proc. of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Seoul, South Korea, 280–289.

[65] Khiet P Truong, Ronald Poppe, Iwan de Kok, and Dirk Heylen. 2011. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Proc. Interspeech*. 4.

[66] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proc. of the International Workshop on Audio/Visual Emotion Challenge*. 3–10. https://doi.org/10.1145/2988257.2988258

[67] Alessandro Vinciarelli, Alfred Dielmann, Sarah Favre, and Hugues Salamin. 2009. Canal9: A database of political debates for analysis of social interactions. In *Proc. of the IEEE International Conference on Affective Computing and Intelligent Interaction and Workshops*. 1–4. https://doi.org/10.1109/ACII.2009.5349466