Graph-based Group Modelling for Backchannel Detection

Garima Sharma
garima.sharma1@monash.edu
Monash University

Kalin Stefanov
kalin.stefanov@monash.edu
Monash University

Abhinav Dhall
abhinav@iitrpr.ac.in
Indian Institute of Technology Ropar
Monash University

Jianfei Cai
jianfei.cai@monash.edu
Monash University

ABSTRACT

The brief responses given by listeners in group conversations are known as backchannels rendering the task of backchannel detection an essential facet of group interaction analysis. Most of the current backchannel detection studies explore various audio-visual cues for individuals. However, analysing all group members is of utmost importance for backchannel detection, like any group interaction. This study uses a graph neural network to model group interaction through all members' implicit and explicit behaviours. The proposed method achieves the best and second best performance on agreement estimation and backchannel detection tasks, respectively, of the 2022 *MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation* challenge.

CCS CONCEPTS

• **Computing methodologies** → Neural networks; Supervised learning; • **Human-centered computing** → Collaborative interaction.

KEYWORDS

graph neural networks; backchannel detection; agreement estimation

ACM Reference Format:
Garima Sharma, Kalin Stefanov, Abhinav Dhall, and Jianfei Cai. 2022. Graph-based Group Modelling for Backchannel Detection. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3503161.3551605

1 INTRODUCTION

The quality of group interactions depends on several factors such as engagement and encouragement [14]. In addition to the speakers', other participants' verbal and non-verbal behaviours are equally important. Backchannels are brief responses given by listeners to show their agreement or assessment in a conversation [21] and their purpose is not to mark interruptions for taking the floor for speaking [12]. Backchannels include words such as *'hmm'*, *'yeah'*, *'okay'*, *'oh!'* and gestures such as head nods and facial expressions.

Backchannel detection has a wide range of real-world applications [11, 15]. Such methods can help robots for better human-robot interaction [5]. Providing backchannels also improves engagement and feedback [11]. Hence, such methods can help indicate a group's cohesion and attentiveness in different contexts. However, many challenges are associated with automatic backchannel detection; for example, the signals used to provide backchannels are highly culture-specific and subjective [8]. Furthermore, for an automated method, it is difficult to quantify the group interaction and separate the audio-visual signals of each group member using data obtained in unconstrained environments.

This work investigates strategies to represent group interaction in terms of graphs. This type of modelling is useful for different tasks, including backchannel detection and agreement estimation. The underlying hypothesis of using graphs to model group interaction is that multiple signals from all group members trigger a person to display a specific behaviour. The study proposes a graph neural network that learns relationships between group members in interaction and achieves the best and second best performance on agreement estimation and backchannel detection tasks, respectively, of the 2022 *MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation* challenge, demonstrating the usefulness of graphs as representations for group interactions.

2 LITERATURE REVIEW

The importance of backchannels in a conversation has been studied for more than five decades. Different studies have discussed factors related to backchannels which are helpful for analysis - both audio and visual backchannels are essential and equivalent if present [12]. Moreover, audio and visual backchannels can replace each other, which is subjective [9]. Since verbal backchannels (and bimodal) might interrupt the speaker, these are often provided at the pauses or the end of sentences. On the other hand, visual backchannels can also be shown in the middle of the sentences [7]. There exist a positive correlation between the number of pauses, mutual gaze exchange and pitch around backchannels with the number of backchannels provided by the listeners [7].

Methods for backchannel prediction are mostly based on language and speech modality. The study of Truong et al. [22] focused on the speaker's pitch and pauses for backchannel prediction. The analysis of the study is termed the pitch and pause model, which shows that backchannel opportunity lies in the pause after the

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3551605

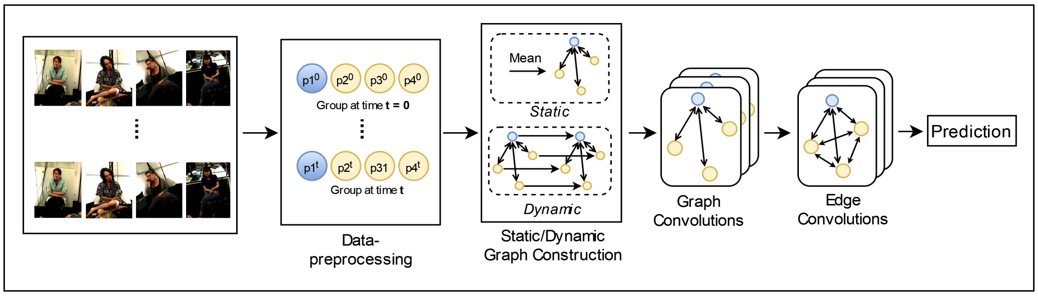

Figure 1: The proposed graph-based method for modelling group interactions. The blue and yellow nodes indicate the main and context members in the group, respectively. *p1-p4 denote each group member.*

speaker's utterance. Morency et al. [15] proposed a probabilistic approach for listener backchannel prediction. The study used multiple features corresponding to prosodic, lexical and visual information. Jain et al. [10] proposed a semi-supervised learning approach for predicting listener's backchannels. Here, acoustic and visual backchannels are separated by training different models using preselected features.

This work proposes a method for group analysis based on graphs. Graph neural networks have gained popularity in problems representing non-Euclidean data with different structures [23] and provides an opportunity to define the structure of the graphs in terms of nodes and edges. Graph neural networks are successfully used in modelling group interactions, such as in social relation analysis [13] and active speaker detection [1, 20]. The literature indicates that head gestures, gaze and speech help to analyse backchannels and most of the existing studies are based on using these signals for listeners or speakers in isolation. However, the implicit relationships between group members are not fully utilised motivating the approach proposed in this work.

3 METHOD

This section defines the problem of backchannel detection and agreement estimation. Two different approaches to analysing the interaction are discussed. First, backchannel detection is done for a given person using only information obtained from that person. Second, data from the whole group is used to detect the backchannel for a given person. These approaches are termed as *individual* and *group* modelling, respectively.

Problem Formulation. Given a t-seconds long video V and audio A of P people interacting, where the video of each person $p_i \in P$ captures the whole face and body of p_i, and the audio contains the speech of all group members. The task is to assign the binary backchannel label $\{0|1\}$ and agreement score in the interval $[-1, +1]$, for a given person. The backchannel label indicates whether the person is providing any backchannel, whereas the agreement score indicates the level of agreement the person displays with the provided backchannel.

Individual Modelling. People use different audio-visual signals to provide backchannels in conversations. The literature review provides information on the relevant signals for the backchannel detection and agreement estimation tasks. For *individual* modelling, the following features are analysed with a support vector machine classifier and regressor with radial basis function kernel for backchannel detection and agreement estimation tasks, respectively. For *individual* modelling, temporal features are combined to create an average representation:

- *Action Units* are indicators used for facial muscle movements. The overall person's facial expression changes during backchannels; hence, action units are functional while analysing backchannels [10].
- *Gaze* exchange is essential to any group conversation. There is a positive correlation between speaker-to-listener gaze exchange and chances of backchannels [7, 15].
- *Head Pose* signals a person's overall visual attention and is useful for contexts in which gaze estimation is difficult.
- *Body Pose* can be helpful in analysing speakers and listeners behaviours and backchannels.
- *Head Gestures* such as head nods are commonly used to signal attention and agreement [7]. These gestures are visual backchannels that listeners can provide at any time.

Group Modelling. The group is modelled as a graph where each group member is represented as a node, and edges define the relationship between the group members. Such a graphical representation can leverage the signals of each group member and the implicit relation between them. Figure 1 illustrates the proposed graph-based *group* modelling. The proposed structure assumes that backchannel labels and agreement scores are present only for one person in the group. This person is marked as the main person and is indicated with blue colour. For simplicity, the label/score of the main person is assigned to all group members. Hence, the model uses information from all group members to detect/estimate the main person's backchannel/agreement.

The graph is constructed such that the main person for whom the prediction needs to be made is connected to all other group members with bidirectional edges. The graph can be created in two ways based on temporal information. In the *static* graph, the features corresponding to each node are the average features across time, resulting in one graph per data sample. In the *dynamic* graph, along with the bidirectional edges between the main and other members, each node is also connected with the same node at the next timestep. After graph construction, graph convolutions are used to learn the node features and the local structure. Further, edge convolutions are used to learn the edge features corresponding to a node and its neighbours based on the edges in the graph.

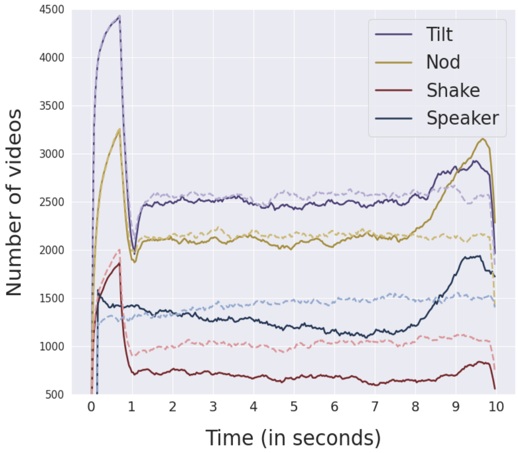

Figure 2: The number of times the main person displays head nods, tilts, shakes and speaks across time. Here, *continuous* and *dashed* lines represent people providing backchannels and not providing backchannels, respectively. The increase in the count indicates that people provide backchannels in the last 1 second in the MPIIInteraction dataset. *Please note that the results before the 1-sec mark can be ignored as the used temporal models require data to stabilise.*

4 EXPERIMENTAL DETAILS

Dataset. We used the MPIIInteraction dataset for backchannel detection and agreement estimation made available by the University of Stuttgart, Germany [16, 17]. The dataset consists of 14468 videos captured in a quiet office. Each interaction video is 10-seconds long and composed of three or four participant conversations in German. The dataset includes a video recording of each group member captured from a camera behind the opposite sitting participants. The recorded audio consists of mixed audio of all participants. An expert annotated each recorded video for the occurrence of backchannel behaviours using holistic perception rather than focusing on a fixed cue. The videos are labelled with a backchannel if the participant provides a backchannel at the end. For the backchannel detection task, the dataset provides labels for 6716, 2854 and 4898 videos for training, validation, and testing, respectively. For the agreement estimation task, the dataset provides scores for 3358, 1427 and 4898 videos for training, validation and testing, respectively.

We analysed the dataset for different audible and visual cues for backchannel detection. As reviewed from the literature, listeners often provide backchannels. We calculated the frame-level score from the PerfectMatch [6] model for all group members to distinguish the speakers from listeners. Specific cues such as head nods display a backchannel and agreement in the visual modality. We used the head gesture detector implemented in the OpenSense [19] library to calculate head nods, shakes and tilts across all video frames. Figure 2 illustrates the results for speaker and head gesture detection. Here, the X-axis represents the time in seconds, and the Y-axis indicates the number of videos. For data samples that include backchannels, there is an increased number of head nods, tilts and shakes towards the end of the video. Similarly, there is an increase in the number of speakers towards the end of videos that include backchannels. The plot indicates that participants mostly provide backchannels in the last 1 second of the videos.

Experimental Settings. Given the analysis shown in Figure 2 and empirical results obtained from data of the last 1, 1.5 and 2 seconds, we used only the last 1 second of the video for backchannel detection and agreement estimation. For action units, gaze and head pose, we used OpenFace [2] library. Body pose features are extracted from OpenPose [4] library. We calculated the head gestures using OpenSense [19] library, which gives the head nods, tilts and shakes. In the case of *individual* and *static group* modelling, we used the features' mean of the absolute differences of adjacent frames from the last 1 second [3, 18]. With these features, the resulting graph had 4 nodes, where the main person is connected to the other group members with bidirectional edges. In addition to these features, the *static* and *dynamic group* models are evaluated using features extracted from pre-trained PerfectMatch [6] model. The PerfectMatch model is used to extract the syncronised audio-visual features for each person and in this case, the graph resulted in having 8 nodes, where the main person's audio and video are connected to all other nodes with bidirectional edges. The graph models include 3 graph convolutions followed by 3 edge convolutions and are trained using Adam optimiser with 0.03 learning rate. The performance is computed in terms of accuracy for backchannel detection and mean squared error for agreement estimation.

5 RESULTS

Baseline. The proposed *individual* and *group* modelling are compared with the challenge baseline method [18], which uses features consisting of the mean of the absolute differences of adjacent frames of action units, head pose, gaze and body pose. Support vector machine classifier and regressor are used with radial basis function kernel for backchannel detection and agreement estimation task, respectively. The results of the challenge baseline method are shown in the first two rows of Table 1.

Individual Modelling. Table 1 provides the results for *individual* modelling. The key observations are:

- Head gestures in terms of head nods, tilts and shakes contribute significantly to backchannel detection, achieving 62% accuracy on the validation set. For agreement estimation, head gestures achieved 0.073 mean squared error on the validation set, which is better than the baseline method. These results show the importance of head gestures in terms of head nods, tilts and shakes for analysis of backchannels. It is to be noted that *HeadPose* features provide the relative head yaw, pitch and roll angles; however, the *HeadGesture* features explicitly correspond to the head gestures.
- The addition of action units, head pose, and gaze to head gestures helps in learning complementary information for backchannel detection and agreement estimation. Among these features, the body pose contributes less in backchannel detection and more in agreement estimation. High validation mean squared error for the agreement estimation task is due to the imbalanced data splits compared to the backchannel detection task, where the data splits are balanced.

Table 1: Backchannel detection and agreement estimation results with support vector machine *individual* modelling.

Features	Backchannel (ACC)	Agreement (MSE)
Validation Set		
AU+HeadPose+Gaze [18]	62.10	0.079
AU+HeadPose+Gaze+ BodyPose [18]	63.90	0.079
HeadGesture	62.00	**0.073**
AU+HeadPose+Gaze+ HeadGesture	67.00	0.081
AU+HeadPose+Gaze+ HeadGesture+BodyPose	**68.00**	0.078
Test Set		
AU+HeadPose+Gaze+ BodyPose [18]	59.60	0.066
AU+HeadPose+Gaze+ HeadGesture+BodyPose	**61.65**	**0.062**

Group Modelling. Table 2 provides the results for *group* modelling. The key observations are:

- For *static group* modelling, the trend in performance is similar to the *individual* modelling for backchannel detection. Here, the use of deep features representing the audio-visual synchronisation for each person is found to be helpful for backchannel detection. Deep features also performed better for the agreement estimation task.
- For *dynamic group* modelling, the used feature sets perform worse than the *static group* modelling. A possible reason can be that the mean features across adjacent frames in the last 1 second are more discriminative than the difference of features across adjacent frames in the *dynamic group* modelling. This result suggests the possibility of improving the proposed *dynamic group* modelling to include temporal changes.
- Except for deep features, *group* modelling achieves slightly lower performance than *individual* modelling. This is because the proposed *group* modelling uses the features of all group members while a label/score is present for only one group member. There is a broad scope for performance improvement of the *group* modelling given the labels/scores for all group members are available or by using any other limited supervision training technique.
- From the results in Table 1 and Table 2, it is clear that the proposed graph-based *group* modelling is able to represent the whole group interaction. By using the features of all group members and limited labels, the *group* modelling performs better than *individual* modelling on the backchannel detection task.

To further validate the importance of graph-based *group* modelling, we used all group members' *AU+HeadPose+Gaze* features and trained a support vector classifier for backchannel detection. The experiment resulted in 62% validation accuracy, worse than when a support vector classifier is trained with only the main person's features. This result demonstrates that graph-based *group* modelling helps to learn relations between all group members.

Table 2: Backchannel detection and agreement estimation results with graph-based *group* modelling. * *denotes results obtained after the official evaluation deadline.*

	Features	Backchannel (ACC)	Agreement (MSE)
Static	*Validation Set*		
	AU+HeadPose+Gaze+ HeadGesture	67.13	0.248
	AU+HeadPose+Gaze+ HeadGesture+BodyPose	67.62	0.245
	DeepFeatures	**69.13**	**0.083**
	Test Set		
	AU+HeadPose+Gaze+ HeadGesture	61.26	0.069*
	AU+HeadPose+Gaze+ HeadGesture+BodyPose	62.10	**0.066***
	DeepFeatures	**62.98***	0.068*
Dynamic	*Validation Set*		
	AU+HeadPose+Gaze+ HeadGesture	65.44	0.083
	DeepFeatures	**67.51**	**0.075**
	Test Set		
	DeepFeatures	**59.96**	**0.067***

6 CONCLUSION

This study proposes a graph-based representation for group interaction modelling. This group representation is validated for backchannel detection and agreement estimation tasks. The proposed method achieves the best and second best performance on agreement estimation and backchannel detection tasks, respectively, of the 2022 *MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation* challenge. However, the proposed method has some limitations which will be addressed in the future. The proposed *group* modelling uses the same label/score for all group members because the label/score is present for only one person in each group. A semi-supervised approach can be used to assign the labels/scores for all group members after learning from one person from each group. Another limitation is related to the fact that the proposed methods do not account for the relation between backchannel detection and agreement estimation. This can be addressed by reformulating the problem as a multi-task learning task. Further, speech signals are highly beneficial for backchannel detection. However, the format of the used dataset and the fact that speech separation is still challenging prevented the inclusion of speech-related features in this study. Addressing these limitations will improve the performance of the proposed graph-based representation for backchannel detection and agreement estimation. This graph-based group interaction modelling can be beneficial for other group analysis tasks.

REFERENCES

[1] Juan León Alcázar, Fabian Caba Heilbron, Ali K. Thabet, and Bernard Ghanem. 2021. MAAS: Multi-modal Assignation for Active Speaker Detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 265–274. https://doi.org/10.1109/ICCV48922.2021.00033

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*. IEEE Computer Society, 59–66. https://doi.org/10.1109/FG.2018.00019

[3] Cigdem Beyan, Vasiliki-Maria Katsageorgiou, and Vittorio Murino. 2017. Moving as a Leader: Detecting Emergent Leadership in Small Groups using Body Pose. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. ACM, 1425–1433. https://doi.org/10.1145/3123266.3123404

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 1302–1310. https://doi.org/10.1109/CVPR.2017.143

[5] Eugene Cho, Nasim Motalebi, S. Shyam Sundar, and Saeed Abdullah. 2022. Alexa as an Active Listener: How Backchanneling Can Elicit Self-Disclosure and Promote User Experience. *CoRR* abs/2204.10191 (2022). https://doi.org/10.48550/arXiv.2204.10191 arXiv:2204.10191

[6] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. 2019. Perfect Match: Improved Cross-modal Embeddings for Audio-visual Synchronisation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 3965–3969. https://doi.org/10.1109/ICASSP.2019.8682524

[7] Gaëlle Ferré and Suzanne Renaudier. 2017. Unimodal and Bimodal Backchannels in Conversational English. In *SEMDIAL 2017*. 27–37.

[8] Bettina Heinz. 2003. Backchannel Responses as Strategic Responses in Bilingual Speakers' Conversations. *Journal of Pragmatics* 35, 7 (2003), 1113–1142.

[9] Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. 2013. Backchannel Relevance Spaces. In *Nordic Prosody XI, Tartu, Estonia, 15-17 August, 2012*. Peter Lang Publishing Group, 137–146.

[10] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring Semi-Supervised Learning for Predicting Listener Backchannels. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*. ACM, 395:1–395:12. https://doi.org/10.1145/3411764.3445449

[11] Laura D Kassner and Kate M Cassada. 2017. Chat it up: Backchanneling to promote reflective practice among in-service teachers. *Journal of Digital Learning in Teacher Education* 33, 4 (2017), 160–168.

[12] Robert M Krauss, Connie M Garlock, Peter D Bricker, and Lee E McMahon. 1977. The Role of Audible and Visible Back-channel Responses in Interpersonal Communication. *Journal of Personality and Social Psychology* 35, 7 (1977), 523.

[13] Wanhua Li, Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. 2020. Graph-Based Social Relation Reasoning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV (Lecture Notes in Computer Science, Vol. 12360)*. Springer, 18–34. https://doi.org/10.1007/978-3-030-58555-6_2

[14] Deepti Mishra, Gonca Gokce Menekse Dalveren, Frode S Volden, and Carly Grace Allen. 2021. Group Discussion in a Blended Environment in Engineering Education. (2021).

[15] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2008. Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In *Intelligent Virtual Agents, 8th International Conference, IVA 2008, Tokyo, Japan, September 1-3, 2008. Proceedings (Lecture Notes in Computer Science, Vol. 5208)*. Springer, 176–190. https://doi.org/10.1007/978-3-540-85483-8_18

[16] 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 *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*. ACM, 4878–4882. https://doi.org/10.1145/3474085.3479219

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

[18] 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 International Conference on Multimedia 2022, Lisboa, Portugal, October 10–14, 2022*. ACM. https://doi.org/10.1145/3503161.3551589

[19] Kalin Stefanov, Baiyu Huang, Zongjian Li, and Mohammad Soleymani. 2020. OpenSense: A Platform for Multimodal Data Acquisition and Behavior Perception. In *ICMI '20: International Conference on Multimodal Interaction, Virtual Event, The Netherlands, October 25-29, 2020*. ACM, 660–664. https://doi.org/10.1145/3382507.3418832

[20] Sydney Thompson, Abhijit Gupta, Anjali W. Gupta, Austin Chen, and Marynel Vázquez. 2021. Conversational Group Detection with Graph Neural Networks. In *ICMI '21: International Conference on Multimodal Interaction, Montréal, QC, Canada, October 18-22, 2021*. ACM, 248–252. https://doi.org/10.1145/3462244.3479963

[21] Jackson Tolins and Jean E Fox Tree. 2014. Addressee backchannels steer narrative development. *Journal of Pragmatics* 70 (2014), 152–164.

[22] Khiet P. Truong, Ronald Poppe, and Dirk Heylen. 2010. A Rule-based Backchannel Prediction Model Using Pitch and Pause Information. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*. ISCA, 3058–3061. http://www.isca-speech.org/archive/interspeech_2010/i10_3058.html

[23] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 32, 1 (2021), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386