Non-verbal Behavior Generation for Virtual Characters in Group Conversations

Ferdinand de Coninck*, Zerrin Yumak*, Guntur Sandino†, Remco Veltkamp*
*Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
Email: mod_ferdi@hotmail.com, z.yumak@uu.nl, r.c.veltkamp@uu.nl
†CleVR B.V.
Delft, The Netherlands Email: mgsandino@clevr.net

***Abstract*—We present an approach to synthesize non-verbal behaviors for virtual characters during group conversations. We employ a probabilistic model and use Dynamic Bayesian Networks to find the correlations between the conversational state and non-verbal behaviors. The parameters of the network are learned by annotating and analyzing the CMU Panoptic dataset. The results are evaluated in comparison to the ground truth data and with user experiments. The behaviors can be generated online and have been integrated with the animation engine of a game company specialized in Virtual Reality applications for Cognitive Behavioral Therapy. To our knowledge, this is the first study that takes into account a data-driven approach to automatically generate non-verbal behaviors during group interactions.**

***Keywords*-group interactions, character animation, gaze and gesture behavior**

I. Introduction

Generating believable animations for virtual characters is essential in virtual environments such as games or virtual reality applications. These environments simulate real world situations such as in bars, malls, schools, offices or other public areas. In the real world, these environments are populated with people doing several activities such as walking or conversing in small groups. In this paper, we focus on generating believable group conversations. In a group conversation, people face towards each other and perform non-verbal behaviors including facial expressions, head and eye movements, gestures and body movements based on the content of the speech and the conversational state.

Producing conversational animations for virtual characters has been a challenging task. While some virtual environment designers generate animations manually, it requires a lot of effort. Rule-based approaches to expressive animation generation [1] [2] have been studied, however these often result in repetitive and unnatural motions [3]. Recent approaches to non-verbal behavior generation largely focus on data-driven approaches [4] [5] [6] [7]. However, these methods focus on individual modalities and animation of a single character. Therefore, they are not suitable for generating animations for a group of virtual characters. Animations during group conversations are multi-party and multi-modal [8]. In other words, the dynamics of the group behavior is important to decide who speaks when with whom and using which non-verbal behaviors. In Cafaro et al. [9], a rule-based method is proposed for the generation of non-verbal behaviors during group interactions. In this paper, we present a data-driven approach by studying the group dynamics and behaviors using real-life data from the CMU Panoptic dataset [10]. We learn the correlations between the conversational state and gaze and gesture behaviors using Dynamic Bayesian Networks [11]. The results show that our approach is able to produce animations automatically for small group conversations of background characters. Figure 1 shows a sample group interaction and behavior components taken into account in our framework. In summary, the contribution of this paper is as follows:

- analysis of gaze and gesture behavior during group conversations using real-life video recordings;
- a Dynamic Bayesian Network that models the correlations between the conversational state and gaze and gesture behaviors;
- quantitative and qualitative evaluation of the model and a novel multi-modal motion synthesis framework for generating group conversations.

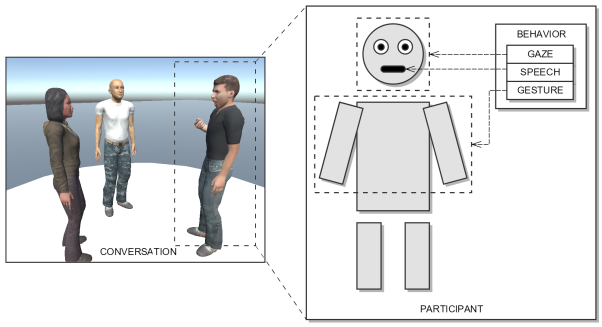

Figure 1. Three-party group interaction and behavior components.

In Section II, we mention the related work with a focus on theoretical background on group interactions and motion synthesis for conversational behavior of virtual characters. Section III presents our approach including the analysis of group interactions and synthesis of gaze and gesture

behavior using Dynamic Bayesian Networks. Evaluation and results are shown in Section IV followed by the conclusions and future work in Section V.

II. Related Work

In this section, we first give a theoretical background on group interactions. Our work is about modeling background characters, thus we are interested in modeling behaviors visible from outside, namely turn-taking, gaze and gesture behaviors. The movement of the eyes and facial expressions are out of the scope. Next we mention the previous works that automatically generate gaze and gesture animation, including the ones on group interactions.

Group interaction involves several participants organized in a formation to interact with each other. Participants position and orient themselves in a way that allows them to address other participants [12]. They take roles such as speaker, addressee, side participant and bystander and coordinate their communicative actions to smoothly interact with other participants. This coordination happens through turn-taking mechanisms. A turn refers to which participant has the right to speak, thus prevents simultaneous speaking. The conversation roles of the participants change over time by taking turns. The speaker can hold the floor (turn holding), hand over the floor (turn yielding) or take over the floor (turn taking) [13]. Gaze is an important tool in conversations and covers several roles. It is used to show the conversation participants who is being addressed or it can indicate someone's intention of holding the speech role or receiving feedback about what someone just said [14]. Non-verbal behaviors during group interactions were analyzed in video recordings [13] [15] [16] and motion capture data [17] but not for the purpose of synthesizing multi-modal behaviors for virtual characters.

Previous works on generating gaze and gesture behavior are based on rule-based or data-driven approaches. For an in-depth view on gaze animation, we refer to the survey by Ruhland et al. [18] which covers both high-level social behaviors and low-level gaze kinematics [19]. Earlier work focused on rule-based approaches [1] [2] [3] by analyzing the text input according to its linguistic and contextual features and creating a rule base to convert these features to appropriate behaviours. The disadvantage of these methods is their inadequacy to explain the full complexity of the mappings between the behaviors and the communicative functions [3]. Data-driven approaches automate this process by finding regularities and dependencies between these factors using statistics, i.e. head movements generation [3] and gesture generation [4]. Levine et. al. [5] used an audio-driven approach and analyzed the pitch and intensity levels in audio and mapped them to the arm movements. Marsella et al. [20] combined text-driven rule-based approaches with an audio-driven approach. More recent approaches employ deep learning for generating non-verbal behaviors [6] [7]. However, these models work for a single character and cannot automatically generate non-verbal behaviors for a group of virtual characters.

There are few previous works that focus on non-verbal behavior synthesis during group interactions. Bohus et al. [21] presented a rule-based gaze model linking turn-taking states to non-verbal behaviors. For example, during a turn hold, the virtual human directed its gaze away from the addressee during the thematic part of the sentence and toward the addressee during the rhematic part of the sentence. In case of multiple participants, it first established eye contact with one addressee and then in turn with each of the other addressees. Wang et al. [22] proposed a rule-based method to model the listening behavior for different roles such as addressee, side participant, overhearer, and eavesdropper. Mutlu et al. [23] studied conversational gaze mechanisms with a humanoid robot based on data collected in a wizard-of-oz setup and developed models of role-signaling, turn-taking, and topic signaling during multi-party interactions. Ennis et. al. [24] [25] conducted user experiments to see how users perceive the non-verbal behaviors of a small groups of agents and found that users are more sensitive to the global movement of the characters compared to specific gesturing behaviors. Thorisson et al. [26] developed a computational model of turn-taking including gaze and manually generated gestures. Cafaro et al. [9] developed a turn-taking state machine and non-verbal behaviors taking into account interpersonal attitudes among the participants based on status and affiliation parameters. Yumak et al. [27] presented an autonomous gaze model to drive the head movement of a virtual receptionist situated in the real-world. However, there is no work that automatically generates gaze and gesture behavior using a data-driven approach for small groups of virtual characters.

III. Our Approach

Our goal is to develop an automated method for the prediction of gaze and gesture behaviors given the conversational state. With conversational state, we refer to the knowledge of who is speaking when and with whom. Other information such as the content of the speech or the audio is not taken into account. Figure 2 shows the steps in the training phase. First, conversation recordings from the data set are converted to annotated conversations. Then, the training data for the gaze and gesture behavior are used to estimate the parameters of the Dynamic Bayesian Network.

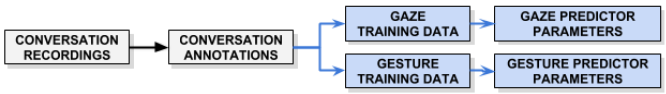

Figure 2. Training phase: Blue parts in the process are automated while the white parts require human input.

We have chosen the CMU Panoptic dataset for this study. There exists other group social interaction data sets such as AMI meeting corpus [28] where people are recorded using video cameras while they are sitting around a meeting table. That was not appropriate for our goals in this paper since we wanted to have scenarios where people are standing in a group formation. We also wanted to have a data set that has Kinect 3D skeletal information for future studies (although skeletal information is not used in this paper). CMU Panoptic data set provides recordings of various conversations such as social games and negotiation scenarios. We have selected three and four-party conversations and annotated the speech, gaze and gesture behavior using the ANVIL annotation tool [29]. More information about the annotated clips are given in Section IV. An example screen shot showing the annotations can be seen in Figure 3. Each participant in the conversation is assigned a unique number. The participant numbers are also used for determining the gaze directions. The speech, gaze and gesture behavior annotation tracks are split for each conversation participant. The speech track defines for each participant the speaking state as either *speaking* or *not speaking*. The gaze direction of a participant is defined as either *looking towards the environment* or *looking towards participant* p. For the labeling of the gestures, the MUMIN Multi-modal Coding Scheme is used [30] and five types of gestures are annotated: *no gesture (idle movement), beat, deictic, metaphoric and iconic*.

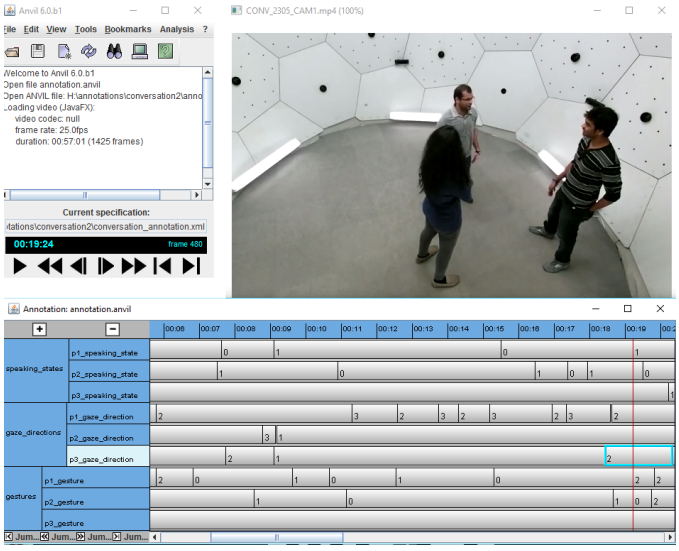

Figure 3. Example annotation in ANVIL.

Inter-coder agreement analysis is performed to validate the annotated labels. A second coder is given the task to annotate a portion of the conversation data set. The similarity of their annotations is measured using Cohen's kappa coefficient. A kappa coefficient between 0.4 and 0.6 is considered fair, a value between 0.6 and 0.75 is considered good and above 0.75 is excellent [31]. Table I shows the kappa coefficients for each participant in the speech, gaze and gesture category. The annotation agreement is rated fair to excellent, with most annotation tracks rated as excellent.

behavior	P1	P2	P3
speech	0.779	0.845	0.861
gaze	0.825	0.860	0.614
gesture	1.0	0.597	0.592

Table I
KAPPA COEFFICIENTS FOR INTER-CODE RELIABILITY ANALYSIS.

The differences in labeling between the coders are discussed after the annotation process. Slight differences in speech behavior annotations are found in the exact timing of starting or ending speech, which is acceptable since this has minimal impact on the model. Some of the short pauses in speech are considered as not-speaking by one coder, while the other coder considers these cases as part of the speech. It is agreed that short pauses are considered as continuing speech. Similar to speech, there are slight differences in the starting times of gaze shifts. Gestures are found to be harder to annotate as it was not very easy to identify the gesture types, especially with ambiguous short and fast gestures. Notice that the kappa coefficient for the gestures of participant 1 is equal to 1.0. That is because participant 1 was using barely any gestures in that annotated sequence.

The turn-taking states in our model is based on the turn-taking state machine defined in Ravenet et al. [32] Instead of their rule-based system, our model learns gaze behaviors from the Panoptic dataset. The turn-taking states are categorized into speaking states (*owning-the-speech* and *competing-for-the-speech*), listening states (*addressed-listener* and *unaddressed-listener*) and transition states (*start-speaking* and *stop-speaking*). The turn-taking states of the participants are induced from the observable behavior in the data set combining the gaze directions and speaking states of the participants. For example, when a participant is an *addressed-listener*, the gaze target is the speaker, whereas in the *unaddressed-listener* state the gaze target might be one of the other participants present in the conversation. Table II shows the logic propositions for determining the turn-taking states. P defines the list of conversation participants, p is the participant the turn-taking state is computed for, q is the other participants and δ is the annotation time interval. To keep the model simple, we didn't define a separate state for interruptions but included these cases as part of the *stop-speaking* state.

A. Analysis of Gaze Behavior

We analyzed the gaze behaviors in three and four-party conversations according to the turn-taking states. The analysis of three-party conversations are explained in this section as an example. Figure 4 shows the probabilities of gaze

turn. state	logic proposition
unaddres.-list.	$\neg p_{speaking}(t) \wedge \forall_{q \in P, q \neq p, q_{speaking}(t)}(q_{gaze}(t) \neq p)$
addres.-list.	$\neg p_{speaking}(t) \wedge \exists_{q \in P, q \neq p, q_{speaking}(t)}(q_{gaze}(t) = p)$
own-speech	$p_{speaking}(t) \wedge \forall_{q \in P, q \neq p}(\neg q_{speaking}(t))$
comp.-speech	$p_{speaking}(t) \wedge \exists_{q \in P, q \neq p}(q_{speaking}(t))$
start-speak.	$p_{speaking}(t) \wedge p_{speaking}(t - \delta)$
stop-speak.	$p_{speaking}(t - \delta) \wedge \neg p_{speaking}(t)$

Table II
LOGIC PROPOSITIONS FOR DETERMINING THE TURN-TAKING STATES.

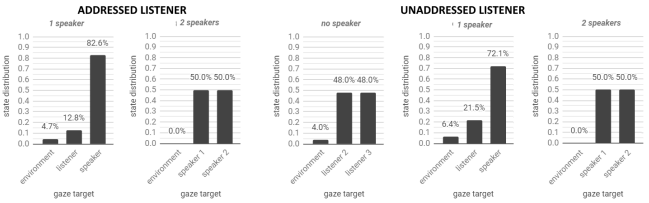

Figure 5. Gaze probabilities for the listening states.

behavior during the speaking turn-taking states (*owning-the-speech* and *competing-for-the-speech*). The participant owning the speech has a gaze target probability of 45.7% towards each listener and 8.6% towards the environment. If the participant is competing for the speech, he/she has a lower probability of 34.3% looking at the listening person because he/she is focused on the other speaker with a probability of 56.9%. When competing for speech against two speakers, the speaker has an equal probability of looking at the speakers (44.1%). The prevalence of mutual gaze interactions between speaking participants and listeners is also analyzed. During 69.89% of the speaking time, the gaze attention is mutual between the speaker and the addressee. This indicates that mutual gaze is a common occurrence and important for the gaze attention of speaking participants.

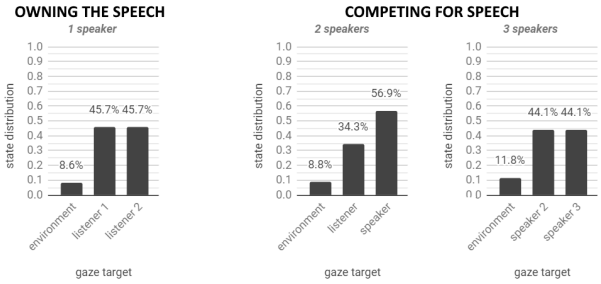

Figure 4. Gaze probabilities for the speaking states.

In the listening state, the participant can be an addressed or an unaddressed listener. Figure 5 shows the probabilities for the listening turn-taking states. If the participant is an addressed listener, he/she gazes towards the speaker 82.6% of the time. The probability drops to 72.1% in the unaddressed state. In both the speaking and listening turn-taking states the participants pay less attention to the environment as their attention is mostly on the speaking person.

Figure 6 shows the gaze behavior during transition states. A participant willing to switch from a listening state to a speaking state (*start-speaking*) has a gaze probability of 62.5% towards the current speaker, 25% towards the listener and the 12.5% towards the environment. When in the *stop-speaking* state, gaze probability towards the speaker is 64.9% and towards the listener is 35.1%. This indicates that the participant gazes most of the time towards the speaking person when switching speaking states. Interesting to note is that the probability for gazing towards the environment when starting speaking is high in comparison to the other states. This is supported by research indicating that participants gaze away when starting speech to avoid feedback from the other participants and focus on the remarks they are about to make [14].

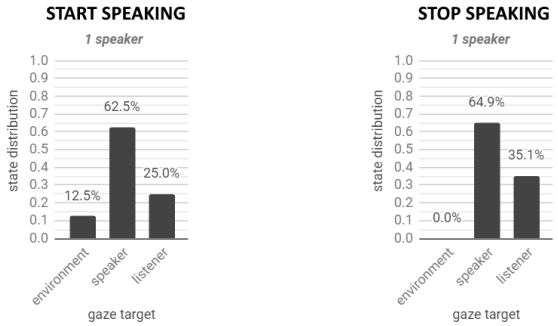

Figure 6. Gaze probabilities for the transition states.

B. Analysis of Gesture Behavior

Table III shows the distribution of gesture behavior according to the turn-taking states. Time spent without performing any gesture is also measured and is defined as being idle. As expected, participants use barely any gestures when listening, with addressed listeners spending 94.16% of the time and unaddressed listeners spending 97.7% of the time being in an idle pose. An interesting observation is that an addressed listener is more likely to use a gesture in comparison to an unaddressed listener. Addressed listeners might be more tempted to give feedback to the speaker and therefore use slightly more gestures, but it is still a rare occurrence. A significant difference in behavior between speaking turn-taking states is that both when starting and ending speech, less gestures are observed in comparison to the other turn-taking states. *Start-speaking* and *stop-speaking* states have an idle gesture percentage of 73.02% and 69.82% respectively. *Own-the-speech* and *compete-for-the-speech* states have a much lower idle gesture percentage being 41.78% and 50.00% respectively. We opted out of delving deeper into the gesture types since they are highly dependent on the style of the conversation. In the data set, for example, a lot of deictic gestures were being used since

the participants were negotiating which resulted in a lot of pointing gestures towards each other. In conclusion, turn-taking states do indeed show differences in gesture frequency and might therefore be used for gesture prediction.

C. Synthesis of Gaze and Gesture Behavior

In this section, we present the gaze and gesture behavior prediction model based on the analysis in the previous section. Humans display a variety of behavior in the same conversation setting. To display this behavior variety, a probabilistic model was chosen. Probabilistic graphical models are suitable for the purpose of modeling behavior given a set of conditions. They provide an intuitive way of defining the parameters of the model and they are able to handle multiple observations of discrete states. Dynamic Bayesian Networks (DBN) [11] were used in this study. DBNs are suitable for online prediction and are efficient since they only require a look-up table representing conditional probabilities between the conversational state and non-verbal behaviors. Another advantage is that the dependencies can be interpreted by a human, which helps to understand how the behaviors are generated and allows for changes if desired. When it is time to predict new gaze or gesture behavior, the relevant variables for the prediction of either gaze or gesture behavior are extracted from the last known conversation state. The behavior prediction is performed on a participant level, meaning that the prediction models are responsible for predicting relevant behavior individually for each participant. For the implementation, the *R* statistical package *bnlearn* [33] is used.

For the prediction of believable gaze behavior, we developed two different models: 1) a prediction model without taking into account the conversational state and 2) a prediction model taking into account the conversational state. The prediction model without conversational state is chosen as the base model and simulates gaze behavior without knowing the behavior of the other participants. It predicts gaze behavior using only the current gaze target of the participant. For the prediction of the gestures, the DBN model takes the turn-taking states as the sole input. The gesture predictor either selects a new gesture type, or no gesture (based on the analysis in Table III).

The DBN model with conversational state is more complex and it will be explained in the next paragraphs. Figure 7 shows the overall process for the prediction of gaze behavior of one participant. For each turn-taking state, a different DBN is defined. Once the appropriate DBN is found (based on the rules defined in Table II), variables from the conversational state are used as input for the selected DBN and gaze target probabilities are generated. The probabilities are used to select a gaze target and then the conversational state is updated with new variables to be used in the next prediction.

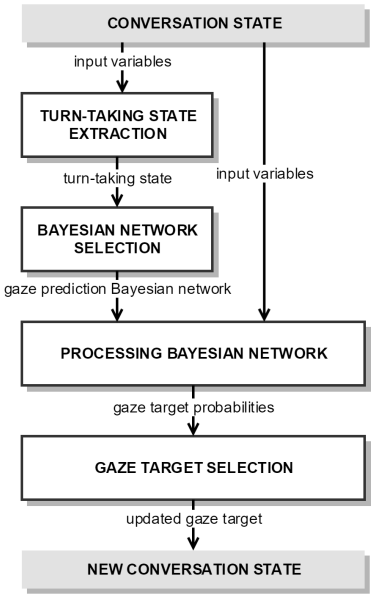

Figure 7. DBN process overview for gaze prediction.

Each turn-taking state has a dedicated Bayesian network determining the variables to be taken into account for that case. When the participant is in the *owning-the-speech* state, the latest known gaze direction is relevant for predicting the next gaze behavior. In other words, it is likely that a person looking at a particular target is likely to look at the same target to have a continuity in the gaze behavior. Another aspect is the mutual gaze based on our analysis in Section IIIA. A participant that owns the speech might have mutual gaze with the other participants who are addressing the speaker. Therefore, the participants addressing the speaker are also taken into account in the DBN model of the *owning-the-speech* state. When in the *competing-for-the-speech* state, the participant should be able to identify the other speakers to act upon accordingly. The speaking states of the other participants is therefore included in the Bayesian network of the *competing-for-the-speech* turn-taking state, in addition to the current gaze target. The listening turn-taking states *addressed-listener* and *unaddressed-listener* require the speaking states of the other participants since listening participants preferably gaze towards a speaker rather than to another listening participant. When in the *addressed-listener* state, the listener also needs to know who is addressing him or her. Between the speaking and listening turn-taking states are the switching turn-taking states: *start-speech* and *end-speech*. When starting speech, the currently speaking participants are of importance. Gaze is directed towards the speakers to indicate intention and request attention. The last speaker should also be taken into account. It may naturally occur that there is a small moment of silence between the

gesture	turn-taking states					
	own speech	compete for speech	addres. listener	unaddres. listener	start speaking	stop speaking
Idle	41.78%	50.00%	94.16%	97.70%	73.02%	69.84%
Beat	31.11%	22.22%	1.95%	0.57%	15.87%	12.70%
Deictic	15.78%	19.84%	0.43%	0.29%	4.76%	6.35%
Iconic	4.89%	3.17%	2.38%	1.15%	3.17%	7.94%
Metap.	6.44%	4.76%	1.08%	0.29%	3.17%	3.17%

Table III
GESTURE BEHAVIOR DURING THE TURN-TAKING STATES.

ending of the speech and the speech reaction of another participant. With only the currently speaking participants as information when starting speech, the participant will not able to respond to the latest speaker. When ending the speech, the most important factors determining the gaze direction are whether the other participants are speaking at that moment or the current gaze direction of the participant, which is often the addressee of the participant recently ended his/her speech.

Figure 8 shows the simulation phase. The behavior predictors use an annotated speech sequence as input, extracts the turn-taking states and predicts the complimentary gaze and gesture movements. The gaze prediction is done every 250 msecs while the gesture prediction is dynamic based on the duration of the current gesture. The predicted states are added to the timeline and the conversation behavior realizer displays the current speech, gaze and gesture behavior. For the realization of actual animations, we use the 3D models and animations generated by a game company specialized in Virtual Reality applications for Cognitive Behavioral Therapy. For the simulation of new conversations, the only manual input required by our pipeline is the annotation of the speech starting and end times. Automatic annotation of speech segments is an interesting option and would result in a fully automated pipeline, however it is out of the scope of this work. Alternative to manual annotation, annotations from example conversation clips may be used as input testing data.

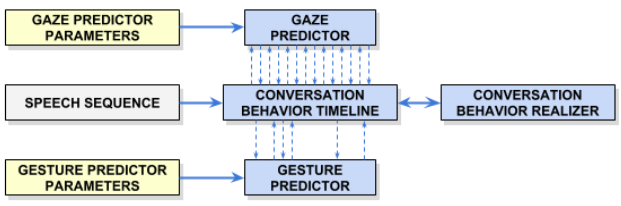

Figure 8. Simulation phase: Blue parts in the process are automated. Yellow parts are generated in the training phase. White parts require human input.

IV. EVALUATION AND RESULTS

The performance of the proposed gaze and gesture prediction models is measured quantitatively and qualitatively. Prediction accuracy of the DBN is used as a quantitative measure. We also conducted a user study to measure the believability of the generated gaze and gesture behaviors in three and four-party conversations.

A. Accuracy of Predictions

The same annotated data set that was used for behavior analysis was used to evaluate the effectiveness of the prediction models. The entire data set contains three three-party conversation clips of 57, 49 and 41 seconds and two four-party conversations of 37 seconds and 54 seconds. By combining the behavior of each participant in a conversation, this data set provides a total of 441 seconds of data for three-party conversations and 364 seconds for the four-party conversations.

For measuring the prediction accuracy, we compared the similarity between predicted non-verbal behaviors and real human behavior in the ground-truth data. The positioning and orientation of the participants were kept same with the original data set. Since the gesture behavior predictor is not using the content of the speech as information, the generated gestures are expected to be far from the original gestures. Therefore the performance of the gesture predictor is only assessed with the user study.

For the gaze behavior, we compared two cases: gaze behavior generated based on the conversational state, and without the conversational state. As described in the previous section, the first one is the full gaze prediction model while the second one is a simplified baseline model that only takes into account the current gaze target to predict the next target. Thus it does not take into account the states of the other participants in the conversation. The initial gaze directions of the participants were set to the initial behavior found in the actual conversations. The prediction models predict the gaze behavior of each participant over the entire duration of the conversation given the speech sequence from the original data set. The prediction accuracy was measured using cross-validation, meaning each conversation is predicted with the remaining conversations being used as

the training data. To avoid a fortunate or unlucky random prediction, the simulations were performed 1000 times for each conversation resulting in a mean prediction accuracy.

The resulting prediction accuracies are displayed in Table IV. Our goal in this study was to find out whether the conversational state has an effect on the generated behaviors in comparison to the base model or randomly generating the behaviors. We found that both gaze prediction model types have an higher accuracy in comparison to random chance which is 33.3% for the three-party conversations and 25% for the four-party conversations (based on the number of possible gaze directions in the conversation). As expected, the prediction model with conversational state has a higher prediction accuracy in comparison to the model without conversational state. This indicates that turn-taking indeed does play a role in determining the gaze direction. However, the low accuracies indicate that the DBN model is not adequate enough to obtain very high accuracy levels with respect to the ground truth data.

	accuracy	
three-party	**without conversat. state**	**with conversat. state**
1	45.5	50.03
2	43.92	48.86
3	45.76	52.18
four-party	**without conversat. state**	**with conversat. state**
1	32.13	39.41
2	31.4	35.85

Table IV
GAZE BEHAVIOR PREDICTION ACCURACY - 3 CLIPS FOR THE THREE-PARTY CONVERSATIONS AND 2 CLIPS FOR THE FOUR-PARTY CONVERSATIONS.

B. User Evaluation

A user study was conducted to measure the believability of the predicted gaze and gesture behavior using an online survey. Our hypothesis was that the gaze behavior with conversational state will have higher believability ratings in comparison to the gaze behavior generated without conversational state. We also expected the gaze and gesture believability ratings to be close to the ground truth believability ratings. The survey was split into two sections which focus on the gaze and gesture prediction separately. A total of 13 participants joined the experiment. Six of them were female and seven were male and they were between the ages of 19 and 27.

For the gaze behavior, 14 clips were generated in total, half of them being three-party and the other half being the four-party conversations. Among the seven three-party conversations, one clip was the ground truth animation. The other six clips were produced based on the gaze prediction model taking the same speech input as the ground truth clip and were generated in two conditions: with or without conversational state. Per condition, three different simulations were generated. The same set-up applies for the seven four-party conversation clips. For the gesture evaluation, four clips were generated for four-party conversations, one of them being the ground-truth animation and the remaining three were predicted using the gesture predictor (three-party conversations were skipped to keep the survey not too long). Each clip were generated focusing on one-modality: For the clips used for the evaluation of the gaze, gestures were taken from the ground-truth data, while for the clips used for the evaluation of the gestures, gaze behaviors from the ground-truth were used. Facial expressions were not included in the simulations but simple mouth movements were added to give an indication of speaking (closing/opening the mouth). Additionally, a red marker is put on top of the head of the speaking participants. The duration of the observed clips were set to 30 seconds to give the participants enough time to make their judgment but also keep the experiment moderate in length. After each conversation clip the participants were asked to rate the believability of the gaze behavior on a scale of 1 to 10 with 1 being unbelievable and 10 being believable. Videos from the user study are submitted as supplementary files.

Table V shows the results of the user study for the gaze prediction. On average the believability of gaze behavior predicted without conversation state is rated the lowest with a rating of 5.31 in three-party conversations and 5.04 in a four-party conversations. As expected, the ground-truth gaze behavior scores the highest with a rating of 6.54 in three-party conversations and 6.42 in four-party conversations. The predicted gaze behavior with conversational state scores in between the ground-truth gaze behavior and gaze behavior predicted without conversational state, with a rating of 5.73 in three-party conversations and 5.9 in four-party conversations. Overall, the scores of the gaze behavior aligns with our initial goals. Table VI shows the believability ratings for gesture prediction. The ratings of the predicted gesture behavior is found to be closer to the ground-truth data, with the original gesture behavior having a rating of 6.96 and the predicted gesture behavior having a rating of 6.41. This shows that it is possible to predict convincing gesture behavior without knowing the content of the speech.

	believability rating	
gaze behavior	**three-party**	**four-party**
predicted without conversational state	5.31 (sd. 1.84)	5.04 (sd. 1.81)
predicted with conversational state	5.73 (sd. 1.62)	5.90 (sd. 1.76)
original gaze behavior	6.54 (sd. 1.63)	6.42 (sd. 1.58)

Table V
BELIEVABILITY RATINGS FOR THE GAZE BEHAVIOR.

V. CONCLUSIONS AND FUTURE WORK

We presented a data-driven method for the generation of non-verbal behaviors during group conversations using

gesture behavior	believability rating
predicted gesture behavior	6.41 (sd. 1.29)
original gesture behavior	6.96 (sd. 1.23)

Table VI
BELIEVABILITY RATINGS FOR THE GESTURE BEHAVIOR.

Dynamic Bayesian Networks. The parameters of the network are learned from a group interaction data set. Our results show that it is possible to generate convincing behaviors for the animation of background characters in group conversations using a data-driven approach. The proposed gaze prediction model taking conversational state into account performs better than the predictor without conversational state. We also found that without knowing the content of the speech, it is possible to predict gestures using the turn-taking states as an indicator. However, there are also limitations to our work and directions of improvement.

First, improvements are required on the model side. Our model does not take into account the initial position and orientation of the participants. That might be one of the causes of low prediction accuracy, since for each conversation to be predicted, the other conversations with a different formation are used as the training data set. Similarly, further characteristics should be taken into account for selecting the training and test sets such as the topic of the conversations (i.e. negotiation, collaboration) or the individual characteristics of the people (i.e. dominant, extrovert). Without taking these variations into account, the model learns an average behavior which might be far from the style of the predicted sequence. Interruptions can also be added as an additional turn-taking state which might improve the believability of the generated behaviors. In our current work, we simplified the model and did not take into account interruption behavior explicitly.

In addition, our work focuses on the high-level modeling of the non-verbal behaviors and the selection of appropriate gaze and gesture actions and do not produce the actual motion trajectories of the joints. Improvements in the animations might lead to better believability. For example, it will be interesting to investigate whether it is possible to automatically generate motion trajectories using speech features and motion capture data. In this work, DBNs were used as a first step to show the feasibility of generating group interactions automatically. However, more complex machine learning models such as deep neural networks and more fine-grained data can be used to achieve better results.

Second, the evaluation of the model and experiment set-up requires improvements. The believability ratings of the ground truth behavior indicates that there is room for improvement in terms of motion quality and rendering. Although the experiment shows promising results, further statistical significance analysis is needed with larger number of participants. The ground-truth simulations also does not take into account the other aspects such as facial expressions. Furthermore, other quantitative evaluation metrics should be defined that represents what a convincing behavior is in that context, e.g. frequency of gaze and gestures, amount of mutual gaze etc.

In conclusion, more complex machine learning models may increase the accuracy of the model and better evaluation metrics should be introduced. Our study is the first one to show the feasibility of generating group conversational animations automatically with a data-driven approach. Finally, we also want to compare our approach with the existing rule-based group interaction models such as [26] and [32].

ACKNOWLEDGMENT

We would like to thank the participants for attending the user study and CleVR for providing the animation engine.

REFERENCES

[1] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: The behavior expression animation toolkit," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '01. New York, NY, USA: ACM, 2001, pp. 477–486. [Online]. Available: http://doi.acm.org/10.1145/383259.383315

[2] D. Heylen, S. Kopp, S. C. Marsella, C. Pelachaud, and H. Vilhjálmsson, "The next step towards a function markup language," in *Proceedings of the 8th International Conference on Intelligent Virtual Agents*, ser. IVA '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 270–280.

[3] J. Lee and S. Marsella, "Modeling speaker behavior: A comparison of two approaches," in *Intelligent Virtual Agents*, Y. Nakano, M. Neff, A. Paiva, and M. Walker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 161–174.

[4] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel, "Gesture modeling and animation based on a probabilistic re-creation of speaker style," *ACM Trans. Graph.*, vol. 27, no. 1, pp. 5:1–5:24, Mar. 2008.

[5] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 124:1–124:11, Jul. 2010. [Online]. Available: http://doi.acm.org/10.1145/1778765.1778861

[6] N. Sadoughi, Y. Liu, and C. Busso, "Meaningful head movements driven by emotional synthetic speech," *Speech Commun.*, vol. 95, no. C, pp. 87–99, Dec. 2017. [Online]. Available: https://doi.org/10.1016/j.specom.2017.07.004

[7] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, ser. IVA '19. New York, NY, USA: ACM, 2019, pp. 97–104. [Online]. Available: http://doi.acm.org/10.1145/3308532.3329472

[8] Z. Yumak and N. Magnenat-Thalmann, *Multimodal and Multi-party Social Interactions*. Cham: Springer International Publishing, 2016, pp. 275–298.

[9] A. Cafaro, B. Ravenet, M. Ochs, H. H. Vilhjálmsson, and C. Pelachaud, "The effects of interpersonal attitude of a group of agents on user's presence and proxemics behavior," *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 2, pp. 12:1–12:33, Jul. 2016. [Online]. Available: http://doi.acm.org/10.1145/2914796

[10] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 3334–3342.

[11] K. P. Murphy, "Dynamic bayesian networks: Representation, inference and learning," Ph.D. dissertation, 2002, aAI3082340.

[12] A. Kendon, "Spacing and orientation in co-present interaction," in *Proceedings of the Second International Conference on Development of Multimodal Interfaces: Active Listening and Synchrony*, ser. COST'09. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 1–15.

[13] K. Jokinen, H. Furukawa, M. Nishida, and S. Yamamoto, "Gaze and turn-taking behavior in casual conversational interactions," *ACM Trans. Interact. Intell. Syst.*, vol. 3, no. 2, pp. 12:1–12:30, Aug. 2013. [Online]. Available: http://doi.acm.org/10.1145/2499474.2499481

[14] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22 – 63, 1967.

[15] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings," *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 1, pp. 4:1–4:31, May 2016. [Online]. Available: http://doi.acm.org/10.1145/2757284

[16] L. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, March 2005.

[17] S. A. Battersby, "Moving together: the organisation of non-verbal cues during multiparty conversation," Ph.D. dissertation, Quenn Mary University of London, 2011.

[18] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell, "A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception," *Comput. Graph. Forum*, vol. 34, no. 6, pp. 299–326, Sep. 2015. [Online]. Available: http://dx.doi.org/10.1111/cgf.12603

[19] T. Pejsa, S. Andrist, M. Gleicher, and B. Mutlu, "Gaze and attention management for embodied conversational agents," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 1, pp. 3:1–3:34, Mar. 2015. [Online]. Available: http://doi.acm.org/10.1145/2724731

[20] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro, "Virtual character performance from speech," in *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '13. New York, NY, USA: ACM, 2013, pp. 25–35. [Online]. Available: http://doi.acm.org/10.1145/2485895.2485900

[21] D. Bohus and E. Horvitz, "Facilitating multiparty dialog with gaze, gesture, and speech," in *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ser. ICMI-MLMI '10. New York, NY, USA: ACM, 2010, pp. 5:1–5:8. [Online]. Available: http://doi.acm.org/10.1145/1891903.1891910

[22] Z. Wang, J. Lee, and S. Marsella, "Multi-party, multi-role comprehensive listening behavior," *Autonomous Agents and Multi-Agent Systems*, vol. 27, pp. 218–234, 2012.

[23] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro, "Conversational gaze mechanisms for humanlike robots," *ACM Trans. Interact. Intell. Syst.*, vol. 1, no. 2, pp. 12:1–12:33, Jan. 2012. [Online]. Available: http://doi.acm.org/10.1145/2070719.2070725

[24] C. Ennis, R. McDonnell, and C. O'Sullivan, "Seeing is believing: Body motion dominates in multisensory conversations," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 91:1–91:9, Jul. 2010. [Online]. Available: http://doi.acm.org/10.1145/1778765.1778828

[25] C. Ennis and C. O'Sullivan, "Perceptually plausible formations for virtual conversers," *Comput. Animat. Virtual Worlds*, vol. 23, no. 3-4, pp. 321–329, May 2012. [Online]. Available: http://dx.doi.org/10.1002/cav.1453

[26] K. R. Thórisson, O. Gislason, G. R. Jonsdottir, and H. T. Thorisson, "A multiparty multimodal architecture for realtime turntaking," in *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, ser. IVA'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 350–356. [Online]. Available: http://dl.acm.org/citation.cfm?id=1889075.1889117

[27] Z. Yumak, B. van den Brink, and A. Egges, "Autonomous social gaze model for an interactive virtual character in real-life settings,," *Computer Animation and Virtual Worlds*, vol. 28, no. 3–4, p. e1757, 2017.

[28] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, ser. MLMI'05. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 28–39.

[29] M. Kipp, "Anvil - a generic annotation tool for multimodal dialogue," in *INTERSPEECH*, 2001.

[30] J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio, "The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena," *Language Resources and Evaluation*, vol. 41, no. 3, pp. 273–287, Dec 2007. [Online]. Available: https://doi.org/10.1007/s10579-007-9061-5

[31] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[32] B. Ravenet, A. Cafaro, B. Biancardi, M. Ochs, and C. Pelachaud, "Conversational behavior reflecting interpersonal attitudes in small group interactions," in *Intelligent Virtual Agents*, W.-P. Brinkman, J. Broekens, and D. Heylen, Eds. Cham: Springer International Publishing, 2015, pp. 375–388.

[33] M. Scutari, "Learning bayesian networks with the bnlearn r package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 2010. [Online]. Available: http://www.jstatsoft.org/v35/i03/