# Development of an Interactive Human/Agent Loop using Multimodal Recurrent Neural Networks

Jieyeon Woo

# Development of an Interactive Human/Agent Loop using Multimodal Recurrent Neural Networks

JIEYEON WOO, ISIR Lab, Sorbonne University, France

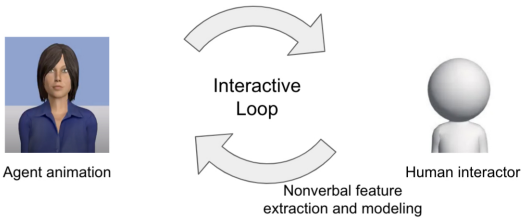


Fig. 1. An interactive loop between a human and an agent modeled through multimodal Recurrent Neural Networks (RNN) for generating facial and head gestures learned from visual and audio features of a dyad

The development of expressive embodied conversational agent (ECA) still remains a big challenge. During an interaction partners continuously adapt their behaviors one to the other [7]. Adaptation mechanisms may take different forms such as the choice of same vocabulary and grammatical form [31], imitation and synchronization [7]. The aim of my PhD project is to improve the interaction between human and agent. The key idea is to create an interactive loop between human and agent which allows the virtual agent to continuously adapt its behavior according to its partner's behavior. The main idea is to learn how dyad of humans adapt their behaviors and implement it into human-agent interaction. My work, based on recurrent neural network, focuses on nonverbal behavior generation and addresses several scientific locks like the multimodality, the intra-personal temporality of multimodal signals or the temporality between partner's social cues. We plan to build a model learned in an end-to-end fashion that generates behaviors considering both acoustic and visual modalities.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Neural networks**.

Additional Key Words and Phrases: Multimodal Interaction, Deep Learning, Embodied Conversational Agent (ECA)

**ACM Reference Format:**
Jieyeon Woo. 2021. Development of an Interactive Human/Agent Loop using Multimodal Recurrent Neural Networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3462244.3481275

## 1 INTRODUCTION

The task of developing expressive embodied conversational agents (ECAs) is a challenging problem. To ensure a fluid and engaging interaction, agents should be able to generate behaviors and react to social signals. To endow an agent with such social capacities requires close investigation on how information is communicated during an interaction. Communication is made of verbal and nonverbal channels. A large part of it is "nonverbal" which refers to "body language" including "gestures, facial expressions, body movement, gaze, dress, and the like to send messages", Burgoon


Manuscript submitted to ACM

et al. [7, p.2]. Generating the behavior of an agent consists therefore not only to generate its words, but also its nonverbal behavior. While conversing, the interlocutors continuously adapt their behavior based on the social signals emitted by their interlocutors. This increases the fluidity of the exchange and the interlocutors' engagement level [19]. Our approach is to analyse interaction partners' behaviors and to propose a computational model to drive the agent's behavior.

In our research, we decide to use Recurrent Neural Networks (RNN) [10] to model the different multimodal nonverbal patterns within dyadic interactions and to focus on generating nonverbal social signals of the virtual agent. For this, it is essential to capture the multimodal signals involved during an interactive loop, the multimodality of the signals, and to ensure their temporal coherence, both intra- and inter-personal temporality.

In this paper, we will start by presenting related work followed up by an explanation of our research, such as our research questions and our approach, our research plan and expected contributions.

## 2 RELATED WORK

### 2.1 Nonverbal signals and their temporality

"Nonverbal behaviors" refer to actions that are distinct from speech such as facial expressions, gestures and postures [7]. It also includes information that may not be explicitly encoded in the verbal behavior such as emotions or social attitudes [3, 33] and thus can be used by one's partner to interpret his/her mental state. Nonverbal signals are also highly related to engagement in the interaction. For example, facial gestures [36] like a smile, gaze direction [28] or posture [15] produced in line with the partner's behaviors can reinforce the interaction. These nonverbal signals are highly multimodal and the timing between them, often referred as "synchrony" [11], is primordial. For an intra-personal point of view, these signals (verb, prosody, gaze, facial gestures, head motion) need to be coordinated with each other. We talk about intra-personal synchrony [6] or intra-personal temporality. Within an interaction, the signals must also be related to nonverbal signals of the partner, which we call inter-personal temporality. As the display of nonverbal behaviors can increase the engagement within the interaction, an agent also need to be able to render these behaviors for an engaging human/agent interaction. Also, the temporal coherence (intra- and inter-personal temporality) must be considered for the interaction.

### 2.2 Nonverbal behavior generation of communicative agents

Behavior generation models for virtual agents can be categorized into two groups: models that focus only on the intra-personal temporality, as only one person is involved, and models that consider the inter-personal temporality, for dyadic interactions.

In contexts where the multi-person interaction is not primitive, approaches focus only on the intra-personal temporality. Some works use simple Feed-Forward Neural Network (FFN) [5] for modality translation to compute communicative behaviors. For example, Karra *et al.* [26] generate 3D facial animation using audio input in real-time. In a similar way, Ding *et al.* [14] propose a FFN regression model to synthesize head motion of a speaker from his/her speech. Other computational methods are also employed. Sadoughi *et al.* [34] propose the use of Bi-directional Long Short-Term Memory (BLSTM) [22] to predict the future head movements from sliding temporal windows of prosody features. A Generative Adversarial Network (GAN) [21] is added to generate multiple realizations of head movements from each speech segment by sampling from a conditioned distribution. Hasegawa *et al.* [23] also use an approach based on BLSTM to predict the 3D human body gesture from audio utterances. Alexanderson *et al.* [2] introduce another

powerful model, based on MoGlow method [24], that generates speech-driven gesticulation. The statistical aspect of the method allows generating several gesticulations from a same speech segment, all with an important plausibility.

Some other works modeled the temporal relationship of nonverbal behaviors between a participant and his/her partner during an interaction (i.e. inter-personal temporality). Earlier behavior generation models for dyadic interactions used rule-based systems such as manually designed rules [35] for backchannel prediction, decision trees [30] for the generation of natural responses and their timing in a chat context or multimodal probabilistic models [27] that are able to predict backchannels using multimodal signals such as prosody, words or gaze. In [18], Feng *et al.* create a FFN model that generates agent's facial gestures based on the agent's and human's facial gestures on previous frames. To the best of our knowledge, it is one of the first works that considers the interactive loop between a user and an embodied agent. Dermouche *et al.* [13] employ LSTM to model the temporality of nonverbal signals and generate ECA's behavior in a dyadic interaction. They introduce the Interactive Loop LSTM (IL-LSTM) that models the agent's nonverbal behaviors by considering both agent's and user's behaviors. In [25], a system that takes audio from both partners and facial expression of human generates corresponding appropriate facial expression of an ECA using an extension of MoGlow [24]. At each time step of the flow, all modalities are encoded using a RNN and their concatenation is passed to a neural network.

### 2.3 Multimodal signal processing

The multimodality of signals that can come from words, prosody, facial expression, … , is an important aspect that needs to be dealt for the task of generating nonverbal behavior to ensure an engaging interaction. The works presented just above use multimodal signals (audio, visual and textual features) for nonverbal behavior generation. Nevertheless, they do not all explicitly model multimodality, thus we observe how these multimodal signals can be processed from models applied for different tasks, including but not limited to nonverbal behavior generation. Chu *et al.* [9] propose a neural conversation model generating facial expression alongside with text. Their goal is to add richness to their generation by exploiting modalities in a separate manner. Rather than concatenating both modalities, they use a RNN dedicated to each modality and then obtain the global description by concatenating the history of each modality. Rajagopalan *et al.* [32] extended the LSTM for multimodal learning by proposing Multi-View LSTM (MV-LSTM) which explicitly models modality-specific and cross-modality interactions. Thus, the model defines four types of memory cells: modality specific cells, coupled cells, fully connected cells and input oriented cells. MV-LSTM shows promising results (high accuracy for the engagement level prediction task) in exploiting multi-view relationships for behavior recognition. Another approach that learns from multiple modalities was proposed by Zadeh *et al.* [38]. Their structure, named Memory Fusion Network (MFN), learns view-specific dynamics in isolation by training a LSTM for each modality and finds cross-view interactions by associating a relevance score to the memory dimensions of each LSTM via an attention mechanism. It stores the cross-view information over time in the Multi-view Gated Memory acting like a dynamic memory module. MFN has been tested on several multimodal databases and show high performance in sentiment analysis, emotion recognition and speaker traits recognition.

Existing models presented above show how we can consider the temporal coherence or explicitly model multimodal signals. Nevertheless, they do not yet fully take into account both aspects of temporality and multimodality for the nonverbal behavior generation. In an interaction both multimodal and temporal relations of exchanged signals can be observed simultaneously and are correlated. The different modalities provide additional information and the capture of complementary information can be strengthened by explicitly modeling multimodal signals. These multimodal signals also need to be temporally coherent with each other. The temporal sync must be ensured not only between the different

modalities of the same person (intra-personal temporality) but also between those of his/her interlocutor (inter-personal temporality). Considering the two aspects together can further help capture and understand the correlation between them. It will thus be interesting to investigate on how to embed their dynamics for an engaging dyadic interaction.

## 3 OUR RESEARCH

We propose to generate nonverbal behaviors of a virtual agent during a dyadic interaction. As stated above, producing nonverbal behavior is important for an ECA as a lot of signals passe through this canal of communication. It requires the generation of a great number of multi-dimensional signals like facial expression (or gesture), head motion, body gesture, posture, prosody and so on. Also, the intra-personal and inter-personal temporality of these signals must be ensured. Moreover, we also plan to integrate our work into the Greta platform [29], which is a 3D humanoid agent capable of communicating with a human using verbal and nonverbal channels. Thus, the model should be causal, taking signals from the past to predict or generate future signals, so that the model can be applied in real-time. These constraints imply some challenges in constructing a model which takes into account all the signals together, from all modalities and all partners. Moreover, to ensure a real-time interaction, the nonverbal behavior needs to be generated at each time step, taking into account the past behaviors of both human and agent. Thus, a loop can be developed to solve the problem of having to consider all the different signals at once and to ensure the real-time interaction. Thus, we choose to use the interactive loop as both human and agent continuously adapt their behaviors within the interaction.

Regarding the bibliography and previously indicated requirements, several questions are still open.

We propose to tackle 3 research questions to model the interactive loop between the human and the agent:

**RQ1:** *Which temporal scale should be considered as input and output to avoid output discontinuity?* We will try to avoid output discontinuity, that comes from independent sequence predictions, by considering the input data at each time step and by predicting the output time step by time step guaranteeing a real-time adaptation. The challenge is to obtain continuous output predictions while avoiding poor learning (vanishing gradient).

**RQ2:** *How to embed multimodal signals' dynamics?* It is important to manage the multimodal aspect of social signals for an engaging interaction. It can be done by a simple concatenation of each modality to form the input of the predictor [13, 18], but we plan to better modelize them with a specific combination of modality encodings [32, 38]. The problem of how to manage multimodal signals in an explainable way needs to be addressed.

**RQ3:** *How to manage intra-personal and inter-personal temporality?* We aim to explicitly model intra- and inter-personal temporal dependencies and merge them, for example with a selective attention module [1]. The main issue of this part is how to connect the modeling of temporal dependencies with the multimodality dynamics model studied in RQ2.

We start by extracting features from our database and apply data processing methods to obtain three distinct modalities. Then, we will construct our model which is based on Dermouche's adaptation model of an interactive loop (IL-LSTM) [13] and improve it by taking into account the temporal coherence and the mulitmodality of signals. Our work will be progressively developed in three stages corresponding to each research question. At the end, all these studies will be integrated in a unique model learned in an end-to-end way.

### 3.1 Corpus and feature extraction

We choose to use the NoXi (NOvice eXpert Interaction) [8] database, a corpus of screen-mediated face-to-face interactions. The database offers dyadic interactions between an expert and a novice in a natural setting. In this work, we use only the French part containing 21 dyadic interactions performed by 28 participants (total duration 7h22). Nonverbal behavior features are obtained through feature extraction. For image processing, two feature vectors are extracted at each time step: a head vector ($x_t^{head}$) composed of head rotations around the 3 axes, and a face vector ($x_t^{face}$) composed of the 17 facial Action Units (AUs) [16], using the opensource toolkit OpenFace [4]. For audio signals, an audio feature vector ($x_t^{audio}$), composed of fundamental frequency, loudness, voicing probability and 13 MFCCs coefficients, is extracted at each time step $t$ via the opensource toolkit openSMILE [17] after a denoising phase. These three features vectors $x_t^{head}$, $x_t^{face}$ and $x_t^{audio}$ are considered as three distinct modalities.

### 3.2 Model architecture

Our purpose is thus to generate the nonverbal behavior of an agent and more particularly, to generate its facial gestures $A_t^{face}$ and head motion $A_t^{head}$ at time step $t$ from facial gestures, head motion and audio of both human partner and agent at previous time $t-1$: $H_{0...t-1}^{face}$, $A_{0...t-1}^{face}$, $H_{0...t-1}^{head}$, $A_{0...t-1}^{head}$, $H_{0...t-1}^{audio}$, $A_{0...t-1}^{audio}$.

We plan to begin with a simple model, like the IL-LSTM in [13], where all modalities for both agent and human of the last 20 frames (Dermouche's parameter [13]) are set as input of a LSTM layer. A fully connected layer allows then to predict outputs.

We tackle our first research question *RQ*1 by avoiding important output discontinuities. They are encountered with the method of IL-LSTM as independent predictions are made for each input sequence and as no previous memory is conserved the predictions are not continuous. To avoid them, we change the paradigm by avoiding the use of a sliding window and using "online LSTM" [37] where cells' memories are continuously updated during the whole interaction. Through this process of updating the cells' memories for each instant, the past is encoded in these memory cells and is used to make a new prediction. Moreover, the model takes its predicted values of the previous time step as input for the prediction at the next time step. Another change we propose on the IL-LSTM model is to symmetrize the problem: predicting the behaviors of both partners. This symmetrization is applied during the training phase, which learns from human/human interactions, as the two human partners are involved in the same way in the interaction and the defined problem or context is identical. Thus, rather than predicting just one behavior from the past data, we predict behaviors of both partners to use all the available information in the loss function (MSE for the IL-LSTM) during the training and thus to help the learning step. This new model is illustrated on Figure 2. Nevertheless, this symmetrization is only applied for the training. For the inference, only the behavior of the agent will be predicted. The optimal time step (temporal scale) found during the training phase of our model, in Figure 2, will be applied to the rest of our project.

In a second stage, we want to tackle the multimodality modeling, by answering our second research question *RQ*2 and plan to employ the MFN proposed in [38], to encode each partner modalities. The idea is to encode the nonverbal behavior of each partner using a MFN to obtain two multimodal memory cells that will be concatenated to predict the future behavior of each partner, as illustrated in Figure 3.

Our last purpose concerns a better modeling of the inter-personal interaction (for the moment, the behaviors of both partners are simply concatenated) using a specific model that has to be developed. We plan to do so by investigating on how both intra-personal and inter-personal temporality can be managed in our model, which corresponds to our last research question *RQ*3.

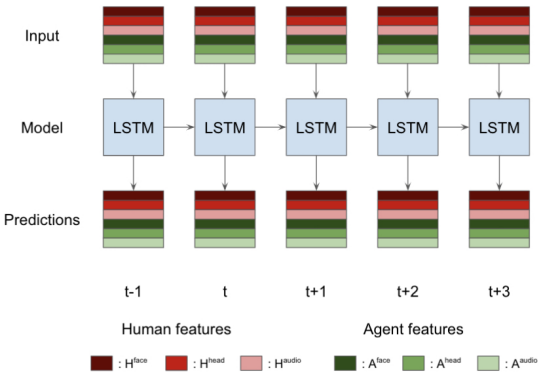


Fig. 2. A first model with a symmetrization for training

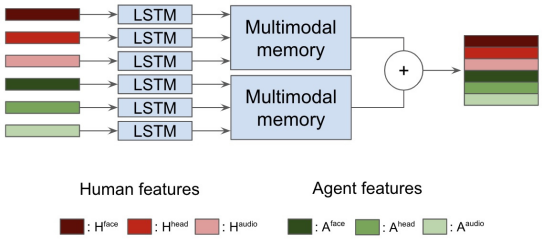


Fig. 3. Fusion of multimodal memory cells

A single end-to-end model will be developed after progressively responding to each of our three research questions.

### 3.3 Evaluation

The task of validating a behavior is very complex. Like any other regression model evaluation, we could accept a generated behavior to be correct based on its quantitative closeness to the ground truth. In other words, the accuracy is calculated using metrics such as MSE [14, 34], RMSE [13], MAE [20], Pearson's correlation [38] for each sample. However, various outcome behaviors could derive from the same surrounding signals. In such cases, where several plausible candidates exist for a given input, quantitative evaluation is not sufficient. Thus, the quality is assessed subjectively via questionnaires [2, 9, 18, 25, 26, 34]. Nevertheless, these conventional methods are not enough to measure the success of our work as we want to evaluate if a behavior, especially nonverbal behavior, is human-like and suitable for a certain instance of a dyadic interaction. For a behavior to be human-like, the motion to which it belongs has to be continuous (i.e. a smooth transition between the previous and the current behavior) and perceived by humans as "natural" (or not as a weird movement). As our work is in a dyadic setting, the synchronization of the predicted behaviors with the corresponding interlocutor's behaviors also needs to be evaluated. Thus, for our project we plan to conduct not only conventional evaluations (both quantitative and qualitative) but also additional studies to evaluate the smoothness and naturalness of the behaviors, and their synchronization with those of its interlocutor. In addition, the IL-LSTM model of Dermouche *et al.* [12], that inspired us in the first place, will be used as baseline, to validate, or not, our propositions.

## 4 RESEARCH PLAN

The research plan and provisional timeline are outlined as the following:

- **2021:** We plan to finalize our first model, Figure 2, for facial gesture and head movement generation of the agent using all input features (visual and audio features) of both agent and human. Then we will address our first research question *RQ*1 by finding the optimal temporal scale for our input and output.
- **2022:** We will develop the multimodal model, Figure 3, and then the inter-personal interaction model to answer the remaining research questions, *RQ*2 and *RQ*3. We intend to develop an end-to-end model and integrate it to the GRETA platform.
- **2023:** We plan to evaluate our model through quantitative and qualitative measures and write the PhD thesis.

## 5 CONTRIBUTIONS

For the PhD project, we expect the following contributions:

(1) Generation of continuous facial gestures and head movements for a virtual agent.
(2) Development of models taking into account multimodal dyamics and intra- and inter-personal temporal dependencies separately.
(3) A single end-to-end model that generates natural nonverbal behaviors by jointly considering multimodal dyamics and intra- and inter-personal temporal dependencies of both agent and human.
(4) Real-time application with the integration into the GRETA platform.

## 6 ACKNOWLEDGMENTS

I would like to thank my PhD advisors, Prof. Catherine Achard and Prof. Catherine Pelachaud, for their support and guidance. This PhD project is performed as a part of IA ANR-DFG-JST Panorama and ANR-JST-CREST TAPAS (19-JSTS-0001-01) projects.

## REFERENCES

[1] Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. 2019. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *2019 International Conference on Multimodal Interaction*. 74–84.
[2] Simon Alexanderson, Gustav Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum* 39 (05 2020), 487–496. https://doi.org/10.1111/cgf.13946
[3] Michael Argyle. 2013. *Bodily communication*. Routledge.
[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.
[5] G. Bebis and M. Georgiopoulos. 1994. Feed-forward neural networks. *IEEE Potentials* 13, 4 (1994), 27–31. https://doi.org/10.1109/45.329294
[6] Carola Bloch, Kai Vogeley, Alexandra L Georgescu, and Christine M Falter-Wagner. 2019. INTRApersonal Synchrony as Constituent of INTERpersonal Synchrony and Its Relevance for Autism Spectrum Disorder. *Frontiers in Robotics and AI* 6 (2019), 73.
[7] Judee K Burgoon, Laura K Guerrero, and Valerie Manusov. 2011. Nonverbal signals. *The SAGE handbook of interpersonal communication* (2011), 239–280.
[8] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth Andre, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. 350–359. https://doi.org/10.1145/3136755.3136780
[9] Hang Chu, D. Li, and S. Fidler. 2018. A Face-to-Face Neural Conversation Model. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 7113–7121.
[10] Axel Cleeremans, David Servan-Schreiber, and James Mcclelland. 1989. Finite State Automata and Simple Recurrent Networks. *Neural Computation - NECO* 1 (09 1989), 372–381. https://doi.org/10.1162/neco.1989.1.3.372
[11] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. 2012. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing* 3, 3 (2012), 349–365.
[12] Soumia Dermouche and Catherine Pelachaud. 2019. Engagement Modeling in Dyadic Interaction. 440–445. https://doi.org/10.1145/3340555.3353765
[13] Soumia Dermouche and Catherine Pelachaud. 2019. Generative model of agent's behaviors in human-agent interaction. In *2019 International Conference on Multimodal Interaction*. 375–384.
[14] Chuang Ding, Lei Xie, and Pengcheng Zhu. 2014. Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications* 74 (07 2014). https://doi.org/10.1007/s11042-014-2156-2
[15] Sidney S D'Mello, Patrick Chipman, and Art Graesser. 2007. Posture as a predictor of learner's affective engagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29.
[16] Paul Ekman and Wallace V Friesen. 1976. Measuring facial movement. *Environmental psychology and nonverbal behavior* 1, 1 (1976), 56–75.
[17] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.
[18] Will Feng, Anitha Kannan, Georgia Gkioxari, and C Lawrence Zitnick. 2017. Learn2Smile: Learning non-verbal interaction through observation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4131–4138.
[19] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3-4 (2003), 143–166.

[20] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 2672–2680.

[22] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5 (2005), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042 IJCNN 2005.

[23] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.

[24] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.

[25] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[26] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

[27] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20, 1 (2010), 70–84.

[28] Yukiko I Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th international conference on Intelligent user interfaces*. 139–148.

[29] Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. 2009. Greta: An interactive expressive ECA system. *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2* 2, 1399–1400. https://doi.org/10.1145/1558109.1558314

[30] Ryota Nishimura, Norihide Kitaoka, and Seiji Nakagawa. 2007. A Spoken Dialog System for Chat-Like Conversations Considering Response Timing, Vol. 4629. 599–606. https://doi.org/10.1007/978-3-540-74628-7_77

[31] Martin Pickering and Simon Garrod. 2004. Toward a Mechanistic Psychology of Dialogue. *The Behavioral and brain sciences* 27 (05 2004), 169–90; discussion 190. https://doi.org/10.1017/S0140525X04000056

[32] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. Extending Long Short-Term Memory for Multi-View Structured Learning, Vol. 9911. 338–353. https://doi.org/10.1007/978-3-319-46478-7_21

[33] Brian Ravenet, Magalie Ochs, and Catherine Pelachaud. 2013. From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In *International workshop on intelligent virtual agents*. Springer, 263–274.

[34] Najmeh Sadoughi and Carlos Busso. 2018. Novel realizations of speech-driven head movements with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6169–6173.

[35] Khiet Truong, Ronald Poppe, and Dirk Heylen. 2010. A rule-based backchannel prediction model using pitch and pause information. *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 3058–3061.

[36] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.

[37] Haimin Yang, Zhisong Pan, and Qing Tao. 2017. Robust and Adaptive Online Time Series Prediction with Long Short-Term Memory. *Computational Intelligence and Neuroscience* 2017 (12 2017), 1–9. https://doi.org/10.1155/2017/9478952

[38] Amir Zadeh, Paul Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning. (02 2018).