Creating an interactive human/agent loop using multimodal recurrent neural networks

Jieyeon Woo, Catherine I Pelachaud, Catherine Achard

Creating an interactive human/agent loop using multimodal recurrent neural networks

Jieyeon Woo
Institut des Systèmes Intelligents et de Robotique, Sorbonne University
Paris, France
woo@isir.upmc.fr

Catherine Pelachaud
CNRS - Institut des Systèmes Intelligents et de Robotique, Sorbonne University
Paris, France
catherine.pelachaud@upmc.fr

Catherine Achard
Institut des Systèmes Intelligents et de Robotique, Sorbonne University
Paris, France
achard@isir.upmc.fr

ABSTRACT

This paper presents a description of ongoing research that aims to improve the interaction between human and Embodied Conversational Agent (ECA). The main idea is to model the interactive loop between human and agent such as the virtual agent can continuously adapt its behavior according to one's partner. This work, based on recurrent neural network, focuses on non-verbal behavior generation and presents several scientific locks like the multimodality, the intra-personal temporality of multimodal signals or the temporality between partner's social cues. The modeling will be done using the NOXI database containing natural human/human interactions and the nonverbal behavior generation will be tested on the GRETA platform that simulates virtual agents.

CCS CONCEPTS

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

KEYWORDS

deep learning, embodied conversational agent (ECA)

1 INTRODUCTION

Embodied Conversational Agents (ECAs) are human-like virtual characters involved in face-to-face interaction with a human or another agent. The goal of researches around ECA is to generate agent behavior and to make them react to verbal and nonverbal social signals, like a human interlocutor. Virtual agents are becoming more and more popular and can be found in applications such as language learning [58], clinical psychology [50], coaching [4] *etc.*

During an interaction, information is communicated in various ways, through verbal and nonverbal channels. The importance of non-verbal communication was demonstrated through many researches. In two experiments, Mehrabian *et al.* [40] concluded that the interpretation of a message is not only through the verbal channel, but also through the vocal and visual channels. A large part of the communication is "nonverbal". But what are these nonverbal signals? In addition to the "body language" including "gestures, facial expressions, body movement, gaze, dress, and the like to send messages", defined by Burgoon et al. [11, p.2], Burgoon *et al.* [11] added some important communicative elements such as use of the voice, touch, distancing and time. Generating the behavior of an agent consists therefore not only to generate its words but also to generate its non-verbal behavior so that it can transmit as much information as a human interlocutor. In this work, we are particularly interesting in this aspect of nonverbal behavior generation.

During a conversation, the interlocutors constantly adapt their behavior based on the social signals emitted by their interlocutors. The behavior adaptation increases the fluidity of the exchange and the interlocutors' engagement level, which are some of the key factors for maintaining a good quality of interaction [25]. As proposed first by Dermouche *et al.* [18], this adaptation can be modeled by an interactive loop between interlocutors that adapt to their surrounding social cues at each iteration. Thus, by studying the other interlocutors' social signals it is possible to produce the appropriate human-like social behavior within an interaction.

In our research, we decide to use Recurrent Neural Networks (RNN) [14] to model this interactive loop. To simplify complex real world scenarios, we chose to work with dyadic interactions and focus on generating nonverbal social signals of the virtual agent. Several scientific locks appear like the multimodality of the signals, the intra-personal temporality between these signals or with partner's social cues. This implies that the multimodal aspect of signals needs to be managed (i.e. multimodality of the signals) and the temporal coherence must be assured between each signal (i.e. intra-personal temporality) for the generation of a natural interaction. The signals also have to be related to verbal and nonverbal signals of the partners (i.e. inter-personal temporality) in order to maintain an interaction with a good quality and an engagement of both partners.

The paper starts with a presentation of the state of the art in Section 2. Then, the database we choose and the considered social signals will be introduced in Section 4.1, before the presentation of the scientific questions posed in Section 3 to implement the interactive loop between human and ECA and some envisaged approaches.

2 RELATED WORK

2.1 Nonverbal signals and their temporality

Terminologically, a "nonverbal behavior" refers to actions that are distinct from speech such as facial expressions, gestures and postures. As stated by Mehrabian [39, p.1], "the term "nonverbal behavior" is a misnomer" as various aspects of speech like fundamental frequency, prosody, … , are traditionally included in the nonverbal phenomena. Even if they are opposed in their denomination, verbal and non-verbal signals reinforce each other. Thus, Bonaccio *et al.* [10] show some examples where nonverbal behavior can repeat verbal discourse (a nod to show agreement), substitute it (an eye roll instead of a statement of contempt), complement it (reddening while talking to an intimidating person), accent it (a slap on the back following a joke) or contradict it (wiping tears away while asserting that one is fine). More importantly, non-verbal behavior can add information that are not explicitly encoded in the verbal behavior such as emotions or social attitudes [5, 52]. Such nonverbal signals are therefore extremely useful to better understand one's partner.

Nonverbal signals are also highly related to engagement in the interaction. For example, facial gestures [57] like a smile, looking at one's partners (gaze direction) [44] or posture [20] can reinforce the interaction. On the opposite, an ECA without facial gestures, with a monotone voice, with a static posture, without any gesture and a perfectly motionless head, will be boring very quickly. Thus, generating nonverbal motion is very important and essential to create a good quality interaction.

These non-verbal signals are highly multimodal and of great dimension. Moreover, the timing between them is primordial. It is often referred as "synchrony" [16]. For an intra-personal point of view, these signals (gaze, facial gestures, posture, body gesture, head motion) need to be coordinated with each other. We talk about intra-personal synchrony [9].

In communication, the Communication Accommodation Theory (CAT) [15] deals with the temporal coherency between speech prosody, speech rate, response latency, laughter or posture. Moreover the inter-personal synchrony is highly correlated with the mimicry effect [33] where, for example, the listener's body moves according to the speaker's rhythm of speech. In [41], Miles *et al.* assert that temporal coordination during dyadic interactions is a foundation for effective social exchange and enhance perceptions of rapport and inter-personal connectedness. Along with synchrony, backchannels [8], which are listener's signals that express attention, interest and understanding through short verbal utterances ("ok"), vocal signals ("uh-huh") or gestural cues (head nod), are also play an important role in maintaining an interaction.

So, displaying nonverbal behavior is important for an ECA. It is required to generate a great number of multi-dimensional signals. They have to be coherent with each other (intra-personal temporality) but also be related to nonverbal signals of the partner (inter-personal temporality).

2.2 Nonverbal behavior generation

We present several works on communicative and interaction behavior generation with a focus on key behaviors that characterize the interaction maintenance or behavior adaptation, such as backchannels and facial gestures. We will first introduce works focusing on single-person based (either speaker or listener) techniques and then for two-person based techniques. These works will be introduced in chronological order.

Earlier studies on interaction behavior generation, notably for ECAs, used rule-based systems. For example, Truong *et al.* [55] manually design rules for backchannel prediction based on pitch and pause information of audio signal. In the same way, Poppe *et al.* [49] evaluate a six rules-based strategy, from the speaker's speech and gaze, for backchannel generation in face-to-face conversations. Decision trees [46] have been employed in a chat context, to generate natural responses and their timing, based on prosodic and surface linguistic information. Morency *et al.* [43] propose a probabilistic model, based on Conditional Random Fields [38], that is able to predict backchannels using multimodal signals such as prosody, words or gaze.

Recent progresses on neural networks and more particularly on recurrent neural networks, combined with the increase of computation capacity, namely Graphics Processor Unit (GPU), have led to a transition from classical approaches to neural ones.

Some works use simple Feed-Forward Neural Network (FFN) [7] for modality translation to compute communicative behaviors. For example, Karra *et al.* [36] generate 3D facial animation using audio input in real time, with low latency. Thus, given a short time window of audio signals, the network infers the facial gestures at the center of the window. A 3D mesh is animated by sliding the window over a vocal audio track. In a similar way, Ding *et al.* [19] proposed FFN regression model to synthesize head motion of a speaker from his/her speech.

New architectures of Neural Networks (NNs) have been constructed to define and update memory cells from previous timesteps. These NNs, called Recurrent Neural Network (RNN) [14] and more especially the Long Short-Term Memory (LSTM) [32], have been broadly used in multiple fields to capture the temporal information.

As stated in Section 2.1, intra-personal and inter-personal temporalities are very important during interactions and thus, need to be properly modeled and employed. So, several works have been interested in using RNN to generate nonverbal behavior.

Always for the purpose of computing nonverbal behavior from a modality, for example from speech to head movement or facial expression, Sadoughi *et al.* [54] propose the use of Bi-directional Long Short-Term Memory (BLSTM) [28] that encode sequences in both directions: forward and backward. This modeling forces to work on sliding temporal windows: prosody features over a time window are used to predict the future head movements. A Generative Adversarial Network (GAN) [27] is also added to generate multiple realizations of head movements from each speech segment by sampling from a conditioned distribution. Hasegawa *et al.* [29] also use an approach based on BLSTM to predict the 3D human body gesture from audio utterances.

GANs are very popular to generate natural sequences and are often used nowadays. For example, temporal GANs with two discriminators have been employed to generate facial gestures from the speech signals of a same person [56], we then talk about "talking head". From a still image of a person and an audio clip containing speech, the model can generate lip movements and natural facial gestures such as blinks and eyebrow movements.

Ginosar *et al.* [26] propose to translate speaker speech into communicative gestures using FFN. As the method suffers from the regression to the mean problem, which leads to overly smooth motion, a GAN has been added. It ensures that generated sequences look like real ones.

Ferstl *et al.* [24] push the use of GAN to map speech to 3D gesture motion by defining several sub-problems, including plausible gesture dynamics, realistic joint configurations, and diverse and smooth motion. Each sub-problem is monitored by separate adversaries.

Alexanderson *et al.* [3] introduce another powerful model, based on MoGlow method [30], 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. Pose sequences are generated via an auto-regressive approach using normalizing flows [48].

The approaches introduced above focus only on the intra-personal temporality as only one person (listener or speaker) is involved. Some other works are more interested in modeling the temporal relationship of nonverbal behaviors between a participant and his/her partner during an interaction.

Huang *et al.* [34] generate facial gestures in human-agent interactions using Conditional Generative Adversarial Networks (CGANs) [42]. At each time step t, a conditional deep convolutional generative adversarial network is employed to produce expressive facial gestures of the interviewer, conditioned on the interviewee's facial gestures. The temporality of signals is not really modeled as previous interviewee's facial gestures descriptors are simply averaged. The main contribution of this work is to generate valid facial expression response thanks to the GAN.

In [23], Feng *et al.* create a Feed-Forward Neural Network (FFN) that generates agent's facial gestures based on the agent's and human's facial gestures on previous frame. To the best of our knowledge, it is one of the first works that considers the interactive loop between user and an embodied agent. The facial descriptors of the 90 last frames of the agent and of the human are concatenated to generate the next 15 frames of agent's facial gestures. As for BLSTM [54], the use of FFN imposes to work on sliding widows.

Dermouche *et al.* [17] employed LSTM to model the temporality of nonverbal signals and generate ECA's behavior in a dyadic interaction. They introduced an Interactive Loop LSTM that models the agent's nonverbal behaviors by considering both agent's and user's behaviors. More precisely, the prediction model takes as input a sequence composed of the last n frames of the agent's and of the user's features and predicts the agent's behavior (smile, head rotation, gaze direction) for the next frame.

In [35], 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 [30]. 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

Previously presented works use multimodal signals (audio, visual and textual features) for nonverbal behavior generation. Nevertheless, they do not study the aspect of multimodality. This multimodality of signals that can come from words, prosody, facial expression, head motion, gestures, … , is an important aspect that needs to be dealt for the task of generating nonverbal behavior.

Chu *et al.* [13] propose a neural conversation model generating facial expression alongside with text. Their goal 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 (facial expression and text). Then, the global description is obtained by concatenating the history of each modality.

Rajagopalan *et al.* [51] 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 in exploiting multi-view relationships for behavior recognition and image caption generation problem. Another approach that learns from multiple modalities was proposed by Zadeh *et al.* [59]. Their structure, named Memory Fusion Network (MFN), learns view-specific dynamics in isolation by training a LSTM for each modality. Then, an attention network is used to find cross-view interactions by associating a relevance score to the memory dimensions of each LSTM. A last component 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.

Another multimodal approach has been proposed to generate animations from natural language sentences [2]. To map linguistic concepts to motion animations, the authors propose a joint embedding of language and pose that is learned end-to-end. The principle is to map sentence and pose to a latent representation using a sentence encoder and a pose encoder respectively. These latent representations should lie close to each other as they represent the same concept. This is ensured during the learning thanks to a specific loss function.

2.4 Evaluation metrics

To validate the quality of a behavior generation model, an evaluation needs to be performed. However, it is difficult to evaluate the quality of behavior generation models as several plausible candidates exist for a given input. For example, one can respond to the interlocutor's smile with smiles of different intensities or timings. Most of the time, evaluations should be conducted both quantitatively and qualitatively.

Concerning the quantitative evaluation, several measures have been employed, according to the type of data to be predicted and to the allowed tolerance around the ground truth.

In some cases, the problem can be transposed to a classification problem with its own metrics. For example, Morency *et al.* [43] propose a method to generate backchannels. They introduce the precision as the probability that a predicted backchannel corresponds to a real one and the recall as the probability that a real backchannel is predicted by the model. The F1-score:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{1}$$

defined as the weighted harmonic mean of precision and recall is also employed.

The same measures have been used in [13] to generate verbal response for example. This part is also evaluated with a F1-score between the words in generated sentences and ground truth.

But in most of the cases, the prediction of nonverbal behavior consists of generating a time-series of real values like position (facial and body landmarks) or magnitude (smile, head rotation, gaze direction, facial action unit). The evaluation then compares the generated time-series $\hat{y}(t)$ to the real one $y(t)$ and the most popular measure is the Mean Square Error (MSE):

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}(t) - y(t))^2 \tag{2}$$

It has for example been used by [19] and [54] to evaluate the synthesized head motion with their models. Similar measures called Mean Absolute Error (MAE) [26]:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\hat{y}(t) - y(t)| \tag{3}$$

Average Position Error (APE) [29]:

$$APE = \frac{1}{T}\sum_{t=1}^{T}\|\hat{y}(t) - y(t)\|_2 \tag{4}$$

or Root Mean Square Error (RMSE) [18]:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}(t) - y(t))^2} \tag{5}$$

have also been used.

Comparison between time-series can also be done using correlation based measures. We can for example cite the Pearson's correlation used in [59]:

$$r = \frac{\sum_{t=1}^{T}(y(t) - \mu_y)(\hat{y}(t) - \mu_{\hat{y}})}{\sqrt{\sum_{t=1}^{T}(y(t) - \mu_y)^2}\sqrt{\sum_{t=1}^{T}(\hat{y}(t) - \mu_{\hat{y}})^2}} \tag{6}$$

or the coefficient of determination (R^2) [18]:

$$r^2 = 1 - \frac{\sum_{t=1}^{T}(y(t) - \hat{y})^2}{\sum_{t=1}^{T}(y(t) - \mu_y)^2} \tag{7}$$

Some particular methods [35, 54] allow the generation of several plausible time-series from a specific input. They can be evaluated by estimating the density of probability $p(\mathbf{x})$ of the generated samples and by computing the likelihood of the test samples using this distribution:

$$L = \prod_i p(\mathbf{x}_i) \tag{8}$$

Some authors [37, 53] consider that the goal of behavior generation is not to reproduce a ground truth behavior but to produce plausible and realistic behavior. Thus, Kucherenko *et al.* [37] propose to study and compare distribution statistics of generated and real gestures and more particularly, statistics on speed and jerk. A similar approach has been adopted in [35] where, in addition to the likelihood of test data, Jonell *et al.* evaluate the rate of change of acceleration and the range of the motion. A statistical comparison between test and real data has also be proposed in [29] that consider naturalness, time constancy and semantic constancy of gesture.

Quantitative evaluation metrics are not sufficient as several plausible candidates exist for a same input and thus quantitative closeness to the ground truth does not signify that the generated behavior is valid. Also, they do not measure how behaviors are perceived by humans nor their impact on the perceived quality of the interaction. To better evaluate behaviors, almost all the works propose qualitative studies that are conducted via questionnaires [3, 13, 23, 35, 36, 54].

3 CHALLENGES

As stated in the Section 2.2, producing nonverbal behavior is important for an ECA as a lot of information passes 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 assured.

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.

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

3.1 Which temporal scale should be considered as input and output?

Considering the input data, two main categories of approaches exist: those that work on temporal moving windows and those that consider each time step as it arrives. The first ones are mainly based on FFN [19, 26, 36] and Bi-directionnal LSTM while the second ones often use RNN or LSTM [3, 17, 29, 54].

Both methods have advantages and drawbacks and thus a comparative study should be interesting. The methods based on a moving window avoid the vanishing gradient problem that often occurs with RNN [31], particularly for long sequence. But, they require to *a priori* set the size of the moving window and, more importantly, can lead to discontinuous results. The opposite conclusion can be drawn for RNN based methods that can be affected by a poor learning (vanishing gradient) but produce more continuous predictions thanks to a memory that is updated with time.

The problem differs for the output: do we have to predict the output time step by time step guaranteeing a real time adaptation to the partner or is it better to make the prediction on “chunk” (i.e. sliding window)? Actually, it seems for example difficult to predict a gesture timestamp by timestamp without deciding the realization of the whole gesture. In the literature, some methods make the prediction at each time step [17, 36], while other ones predict the whole sequence using sequence translation [54]. A compromise has probably to be find but some studies on the quality of the interaction according to this temporal length have to be undertaken.

3.2 How to manage multimodality?

Social cues are by definition highly multimodal and includes signals coming from text (words, dialog act,…), from audio (prosody, turn-talking, interruption,…), from image (facial gesture, posture, body gesture, head motion,…). In addition to being multimodal, these signals have different types like series of numerical values (fundamental frequency, head rotation for example) or series of categorical variables (words, dialog act, turn-taking,...). Thus, they probably have to be processed differently, even if it is important to consider them altogether.

This multimodality has to be managed both for input signals and output signals. Actually, all signals are extracted from the human partner and their modeling requires the management of their multimodal aspect. For the output side, all these signals have to be generated to animate an ECA. Some works consider multimodality just by concatenating each modality to form the input of the predictor [17, 23], but such a solution is probably not optimal as processing multimodal signals in an individual manner can result to a better exploitation. In [35] or [13], each modality is encoded by a specific RNN, for a finer extraction of information of each modality features, and the modality encodings were then concatenated for the final prediction.

The management of multimodal signals is not specific to ECA behavior generation and has been studied for other applications. Some interesting methods have been proposed to better combine modality encodings [51, 59]. They obtain important gains in performance on tasks like sentiment analysis, emotion recognition or speaker traits recognition and it should be interested to use these models for ECA multimodal behavior synthesis. Multi-modal signals modeling recently appears in literature [47] and there is no doubt that a lot of work remains to be done on this subject.

3.3 How to manage intra-personal and inter-personal temporality?

Modeling an interaction between partners is complex, even for a dyadic interaction. Many signals of both partners are involved with a lot of temporal dependency between them. A dependency exists between partners. For example, when one person smiles, the other person often responds to his/her smile. It is a short term dependency. Moreover, individuals often adapt their speech rate or voice level to their partner. This time dependency is long-term, much longer than the first example. But the generation of ECA's nonverbal behavior must not only consider signals from his/her partner but also the temporal dependency of its own signals. Actually, the generated signals have also a strong temporal dependency amongst themselves. For example, head motion or hand gesture can reinforce speech, a change in posture may indicate a desire to engage a new turn talking. In a more obvious manner, lip motion or facial gestures are highly dependent on speech.

But how do we manage all these temporal dependencies, both inter-personal and intra-personal?

In most of the works, this aspect is not addressed and a generic model is used and supposed to do the job. We believe, however, that a more specific model, explicitly modeling intra- and inter-personal dependencies, could help the generation of more natural behaviors. To the best of our knowledge, Ahuja *et al.* [1] were the first and only ones to propose such a model. They explicitly model intra-personal and inter-personal dynamics and merge them using a selective attention module to generate sequences of body poses. Based on this first work, other NN architectures can be proposed and compared.

4 OUR RESEARCH

In this work, we propose to generate nonverbal behaviors of a virtual agent during a dyadic interaction. As we plan to integrate our work in an human-agent interaction platform, particular attention will be paid on real time aspect. Moreover, the model should be causal, taking only information from the past, so that the model can be applied in real-time.

We propose to tackle the 3 challenges mentioned above to model the interacting loop between the human and the agent. This model will be constructed using a dyadic human/human interactions database.

4.1 Corpus

We choose to use the NoXi (NOvice eXpert Interaction) [12] database, a corpus of screen-mediated face-to-face interactions recorded in three countries (France, Germany and UK), spoken in seven languages (English, French, German, Spanish, Indonesian, Arabic and Italian). The screen-mediated recording allowed recording a face-to-face conversation without the use of multiple cameras placed in different angles.

The protocol is constructed to capture full body movements, facial expressions, gestures and speech. They are acquired using the Microsoft's Kinect 2 and a dynamic head-set microphone.

The database offers 25 hours of dyadic interactions in a natural setting. It is obtained by enrolling in each dyadic interaction, an expert willing to share his/her experience and a novice willing to learn on a given topic. In this work, we use only the French part containing 21 dyadic interactions performed by 28 participants (total duration 7h22).

Nonverbal behaviors need to be extracted from the database and an analysis should be done for model training. The features constructing these behaviors were obtained through feature extraction which was done separately for audio and video.

For images processing, we use the opensource toolkit OpenFace [6] on both expert and novice videos. It allows us to estimate head pose and rotation as well as facial Action Units (AUs) that represent the movements of facial muscles classified according to the FACS (Facial Action Coding System) taxonomy [21]. All features have been filtered using a median filter. Moreover, when OpenFace does not succeed to extract AUs, the missing values are estimated using a linear interpolation. Two feature vectors are thus extracted at each time step, one x_t^{head}, composed of head rotations around the 3 axes and head position, and the second one x_t^{face}, composed of the 17 AUs extracted by OpenFace. We split the visual features into two separate vectors (x_t^{head} and x_t^{face}), instead of putting them into a single vector, as head rotations and AUs have distinct feature characteristics.

Audio signals are first filtered to decrease background noise and to eliminate speech from the other interlocutor. Then the opensource toolkit openSMILE [22] is employed to extract speech related features. As for video, a median filter is then applied. The feature vector, extracted at each time step x_t^{audio}, is composed of fundamental frequency, loudness, voicing probability and 13 MFCCs coefficients.

In the following, these three features vectors x_t^{head}, x_t^{face} and x_t^{audio} will be considered as three distinct modalities.

4.2 Our contribution

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: $H^{face}_{0...t-1}$, $A^{face}_{0...t-1}$, $H^{head}_{0...t-1}$, $A^{head}_{0...t-1}$, $H^{audio}_{0...t-1}$, $A^{audio}_{0...t-1}$.

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.

We plan to begin with a simple model, like the one proposed in [17] named IL-LSTM, where all modalities for both agent and human of the last 20 frames are set as input of a LSTM layer. A fully connected layer allows then to predict outputs, like illustrated in Figure 1.

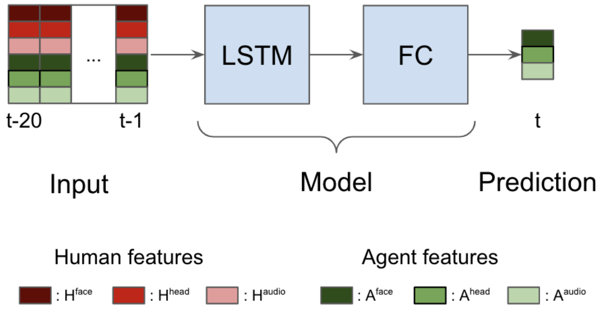

Figure 1: A single LSTM working on a sliding window

To avoid important output discontinuities encountered with this method, the problem which triggers our first research question stated in Section 3.1, we change the paradigm by avoiding the use of a sliding window and using "online LSTM" where cells' memories are continuously updated during the whole interaction. In such a way, the past is encoded in these memory cells and is used to make new prediction. Moreover, the model takes its predicted values of 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: during learning on human/human interactions, both partners are involved in the same way in the interaction. Thus, rather than predicting just one behavior from the past data, we predict both behaviors. This allows us 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.

The IL-LSTM model will be used as baseline, to validate, or not, these first two propositions and particularly to study the use of temporal window.

In a second stage, we want to tackle the multimodality modeling, by answering our second research question in Section 3.2, and plan to employ the Memory Fusion Network (MFN) proposed in [59] to encode each partner modalities. This model, illustrated in Figure 3, encodes independently each modality using a specific LSTM and then, construct a multimodal gated memory using an attention mechanism. Thus, 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 4.

Our last purpose, concerns a better modeling of the inter-personal interaction (for the moment, just a concatenation is employed) using a specific model that has to be developed. We plan to do so by investigating the how both intra-personal and inter-personal temporality can be managed in our model, which corresponds to our last research question in Section 3.3.

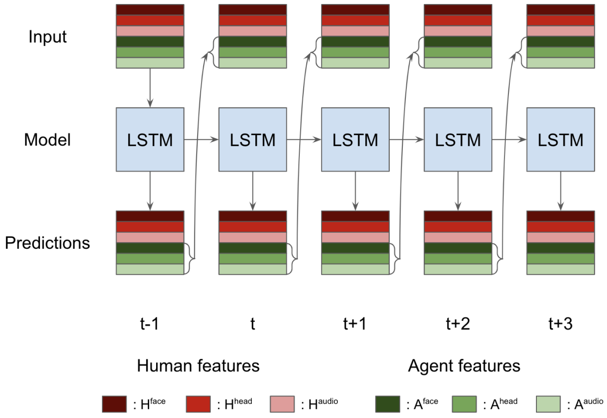

Figure 2: A first model with a symmetrization of the problem

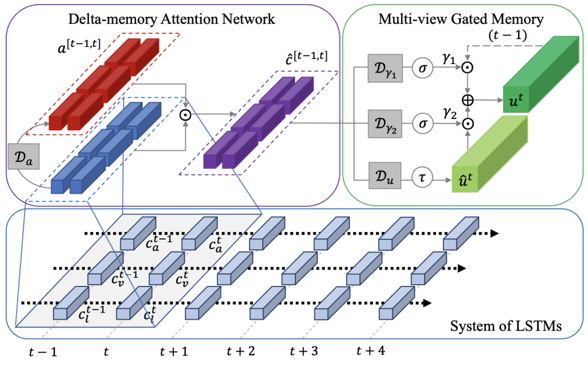

Figure 3: MFN pipeline in [59]

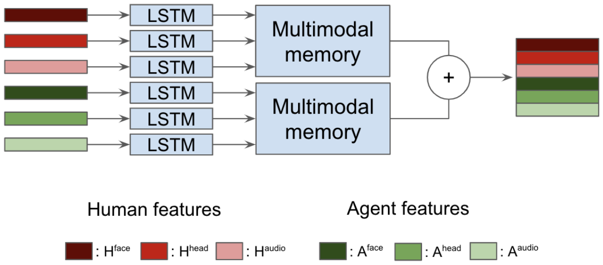

Figure 4: Fusion of multimodal memory cells

4.3 Evaluation

For the evaluation, we plan to evaluate our models through quantitative and qualitative measures as presented in Section 2.4. The baseline will be the model of Dermouche *et al.* [17] that inspired us in the first place.

Moreover, the best model will be integrated on the Greta platform [45], which is a 3D humanoid agent capable of communicating with a human using verbal and nonverbal channels.

5 CONCLUSION

This paper presents a review of the state of the art on nonverbal behavior generation for ECA. Following this review, three remaining scientific locks have been identified. We propose to study them during our future works and propose some perspectives at this end.

6 ACKNOWLEDGMENTS

This work was 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] Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*. IEEE, 719–728.

[3] 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

[4] Keith Anderson, Elisabeth André, Tobias Baur, Sara Bernardini, Mathieu Chollet, Evi Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, et al. 2013. The TARDIS framework: intelligent virtual agents for social coaching in job interviews. In *International Conference on Advances in Computer Entertainment Technology*. Springer, 476–491.

[5] Michael Argyle. 2013. *Bodily communication*. Routledge.

[6] 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.

[7] G. Bebis and M. Georgiopoulos. 1994. Feed-forward neural networks. *IEEE Potentials* 13, 4 (1994), 27–31. https://doi.org/10.1109/45.329294

[8] Roxane Bertrand, Gaëlle Ferré, Philippe Blache, Robert Espesser, and Stéphane Rauzy. 2007. Backchannels revisited from a multimodal perspective.

[9] 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.

[10] Silvia Bonaccio, Jane O'Reilly, Sharon L O'Sullivan, and François Chiocchio. 2016. Nonverbal behavior and communication in the workplace: A review and an agenda for research. *Journal of Management* 42, 5 (2016), 1044–1074.

[11] Judee K Burgoon, Laura K Guerrero, and Valerie Manusov. 2011. Nonverbal signals. *The SAGE handbook of interpersonal communication* (2011), 239–280.

[12] 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

[13] 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.

[14] 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

[15] Justine Coupland, Nikolas Coupland, and Howard Giles. 1991. Accommodation theory. communication, context and consequences. *Contexts of accommodation* (1991), 1–68.

[16] 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.

[17] Soumia Dermouche and Catherine Pelachaud. 2019. Engagement Modeling in Dyadic Interaction. 440–445. https://doi.org/10.1145/3340555.3353765

[18] 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.

[19] 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

[20] 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.

[21] Paul Ekman and Wallace V Friesen. 1976. Measuring facial movement. *Environmental psychology and nonverbal behavior* 1, 1 (1976), 56–75.

[22] 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.

[23] 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.

[24] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*. 1–10.

[25] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3-4 (2003), 143–166.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, 02 (1998), 107–116.

[32] S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[33] Michael J Hove and Jane L Risen. 2009. It's all in the timing: Interpersonal synchrony increases affiliation. *Social cognition* 27, 6 (2009), 949–960.

[34] Yuchi Huang and Saad Khan. 2017. DyadGAN: Generating Facial Expressions in Dyadic Interactions. 2259–2266. https://doi.org/10.1109/CVPRW.2017.280

[35] 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.

[36] 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.

[37] Taras Kucherenko, D. Hasegawa, G. Henter, N. Kaneko, and H. Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven Gesture Generation. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (2019).

[38] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289.

[39] Albert Mehrabian. 1970. A semantic space for nonverbal behavior. *Journal of consulting and clinical Psychology* 35, 2 (1970), 248.

[40] Albert Mehrabian and Morton Wiener. 1967. Decoding of inconsistent communications. *Journal of personality and social psychology* 6, 1 (1967), 109.

[41] Lynden K Miles, Louise K Nind, and C Neil Macrae. 2009. The rhythm of rapport: Interpersonal synchrony and social perception. *Journal of experimental social psychology* 45, 3 (2009), 585–589.

[42] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[43] 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.

[44] 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.

[45] 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

[46] 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

[47] Kay L O'Halloran. 2011. Multimodal discourse analysis. *Continuum companion to discourse analysis* (2011), 120–137.

[48] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2019. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762* (2019).

[49] Ronald Poppe, Khiet P Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel strategies for artificial listeners. In *International Conference on Intelligent Virtual Agents*. Springer, 146–158.

[50] Simon Provoost, Ho Ming Lau, Jeroen Ruwaard, and Heleen Riper. 2017. Embodied conversational agents in clinical psychology: a scoping review. *Journal of medical Internet research* 19, 5 (2017), e151.

[51] 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

[52] 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.

[53] Brandon Rohrer, Susan Fasoli, Hermano Krebs, Richard Hughes, Bruce Volpe, Walter Frontera, Joel Stein, and Neville Hogan. 2002. Movement Smoothness Changes during Stroke Recovery. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 22 (10 2002), 8297–304. https://doi.org/10.1523/JNEUROSCI.22-18-08297.2002

[54] 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.

[55] 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.

[56] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2019. End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs.. In *CVPR Workshops*. 37–40.

[57] 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.

[58] Preben Wik and Anna Hjalmarsson. 2009. Embodied conversational agents in computer assisted language learning. *Speech communication* 51, 10 (2009), 1024–1037.

[59] 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).