ASAP: Endowing Adaptation Capability to Agent in Human-Agent Interaction

Jieyeon Woo, Catherine Pelachaud, Catherine Achard

▶ **To cite this version:**

Jieyeon Woo, Catherine Pelachaud, Catherine Achard. ASAP: Endowing Adaptation Capability to Agent in Human-Agent Interaction. IUI '23: 28th International Conference on Intelligent User Interfaces, Mar 2023, Sydney NSW Australia, Australia. pp.464-475, 10.1145/3581641.3584081 . hal-04293272

HAL Id: hal-04293272
https://hal.science/hal-04293272

Submitted on 21 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ASAP: Endowing Adaptation Capability to Agent in Human-Agent Interaction

Jieyeon Woo
woo@isir.upmc.fr
ISIR - Sorbonne University
Paris, France

Catherine Pelachaud
catherine.pelachaud@upmc.fr
CNRS - ISIR - Sorbonne University
Paris, France

Catherine Achard
achard@isir.upmc.fr
ISIR - Sorbonne University
Paris, France

ABSTRACT

Socially Interactive Agents (SIAs) offer users with interactive face-to-face conversations. They can take the role of a speaker and communicate verbally and nonverbally their intentions and emotional states; but they should also act as active listener and be an interactive partner. In human-human interaction, interlocutors adapt their behaviors reciprocally and dynamically. The endowment of such adaptation capability can allow SIAs to show social and engaging behaviors. In this paper, we focus on modelizing the reciprocal adaptation to generate SIA behaviors for both conversational roles of speaker and listener. We propose the Augmented Self-Attention Pruning (ASAP) neural network model. ASAP incorporates recurrent neural network, attention mechanism of transformers, and pruning technique to learn the reciprocal adaptation via multimodal social signals. We evaluate our work objectively, via several metrics, and subjectively, through a user perception study where the SIA behaviors generated by ASAP is compared with those of other state-of-the-art models. Our results demonstrate that ASAP significantly outperforms the state-of-the-art models and thus shows the importance of reciprocal adaptation modeling.

CCS CONCEPTS

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

KEYWORDS

socially interactive agent (SIA), reciprocal adaptation, multimodal

ACM Reference Format:
Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. 2023. ASAP: Endowing Adaptation Capability to Agent in Human-Agent Interaction. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581641.3584081

1 INTRODUCTION

An increase in application of embodied agents, such as socially interactive agents (SIAs) (also referred as embodied conversational agents (ECAs) or virtual agents) simulated via a graphical user interface (GUI) or robots with a physical body, can be seen in our daily life. The use of embodied agents ranges from providing assistance to being a companion [34, 54]. To carry out interactive and natural conversations, they are often designed with friendly or human-like appearances and communicative capabilities are given. While the process of starting a conversation and interacting with other people comes naturally to us, it is challenging to endow the same capability of communicating thoughts and intentions to SIAs as it involves complex mechanisms such as planning what to say and how, while talking into account its human interlocutor's behavior.

During an interaction, we constantly coordinate our behavior by perceiving and responding to social signals [9]. This behavior coordination happens with a specific temporality and appears between the signals of a same person (intrapersonally, such as the coordination of facial expression, gesture, and prosody) and between the interlocutors (interpersonally, for example when participants mirror each other's behaviors). The interpersonal coordination (or synchrony) is mutual and evolves during the entire interaction [57]. It can also maintain interlocutors' engagement [14, 28]. Due to the mutuality, temporality, and everlasting facade seen during human-human interactions, human interlocutors can adapt their behaviors continuously to those of the others reciprocally and dynamically. We refer to this adaptation as reciprocal adaptation. It arises in real-time following a looped process.

Communication consists of verbal and nonverbal signals [8]. Nonverbal signals, which are also referred to as body language (including gestures, facial expressions, body movement, and gaze), constitute a major part of communication signals. When generating SIA behaviors, the generation of words (i.e. verbal behavior) might be essential for conveying intentions but nonverbal behavior generation is also important for communicating intentions and to be socially interactive. Recent works on multimodal behavior generation (where only one person is concerned) show promising results for the generation of communicative nonverbal behaviors focusing on Deep Learning (DL) techniques from classical Feed-Forward Neural Network (FFN) to latest Transformers model [3, 7, 16, 19, 21, 27, 29, 32, 56, 72]. These works model the communicative behaviors linked to speech but do not pay attention to the social signals arising between interaction participants. In this paper, we focus on generating nonverbal behavior for dyadic interactions. We aim to provide SIAs with this capacity of reciprocal adaptation to enhance its behaviors so that they can behave naturally like a human-being. We use multimodal features (visual and acoustic) and produce SIA behaviors of an active interactant as both listener and speaker. We hold attention to the aspect of behavior coherence, synchrony, and continuity. Behaviors are made up of continuous values which evolve over time (for example for human motion the body landmark positions change smoothly in time). We also intend to assure the production of continuous behaviors by looking at their temporal continuity which motivates us to look

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of a national government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
IUI '23, March 27–31, 2023, Sydney, NSW, Australia
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0106-1/23/03…$15.00
https://doi.org/10.1145/3581641.3584081

into different prediction approaches of offline and online prediction to better model the motion fluidity. Behavior motions should not only be continuous but also coherent and in sync with those shown by the interactant. We are thus interested in the temporal alignment and the appropriateness of the generated SIA behavior type (e.g. a smile in response of an interlocutor's smile). We also look into how the quality (continuity, temporal alignment, and type of behavior) of the generated behaviors could be quantified via objective measures that are used to evaluate the generated behavior sequences.

With the goal to create a SIA capable of adapting its behaviors to its interlocutor, we propose the Augmented Self-Attention Pruning (ASAP) model that models the reciprocal adaptation of interaction partners throughout the interaction. The multimodal signal information of both interaction partners along with the interpersonal relationship between them are captured. Specifically, ASAP allows us to: (1) capture multimodal information of visual and acoustic features; (2) learn from both interactants through data augmentation technique; (3) better select key features within the interaction via the self-attention mechanism with pruning; (4) generate continuous nonverbal behaviors by updating cells' memories at each step of the inference phase with autoregressive adaptive online prediction; (5) generate behaviors as both active listener and speaker; (6) and train without needing a massive amount of data.

Our paper makes the following contributions:

- We propose the modeling of reciprocal adaptation and show how the endowment of such capability can make SIAs behave more social and engaged as both speaker and listener;
- Our results show that ASAP out-performs state-of-the-art models quantitatively and qualitatively notably for interaction synchrony and engagement.

The rest of the paper is structured as the following: Section 2 presents state-of-the-art of related techniques for continuous nonverbal behavior prediction and evaluation measures; Section 3 introduces the database and feature extraction; Section 4 details the implementation of our ASAP model; Section 5 provides objective and subjective evaluation results; Section 6 summarizes our findings; and Section 7 discusses the practical and social implications of our work of endowing SIAs with reciprocal adaptation capability.

2 RELATED WORK

Related works that are key to our interest of generating social and engaging nonverbal behaviors of SIAs and methods of evaluating these behaviors within interactions quantitatively are outlined in this section.

2.1 Sequence prediction techniques

Generating nonverbal behaviors can be considered as a similar problem as forecasting future non-linguistic action sequences. It is thus interesting to investigate existing sequence prediction techniques that could be applicable to nonverbal behaviors.

The methods of sequence prediction can be broadly split into two: offline and online prediction. Offline prediction predicts by giving a sequence data all at once while online prediction refers to the inference method in which data is predicted sequentially one after another.

2.1.1 Offline prediction. Offline prediction infers with the whole input data given from the start. The prediction is done in chunks and is done independently without considering the previously outputted prediction. Its application can be easily seen for sequence to sequence predictions. Models for such predictions generally have the structure of an autoencoder which consists of an encoder that encodes the inputted sequence and a decoder that predicts the resulting sequence by decoding the output of the encoder. Sequence to sequence prediction models produce good results for machine translation [59] and speech recognition [38]. The representative models that can be seen in the literature are Bidirectional Long Short-Term Memory (BLSTM) [26], Conditional Variational Autoencoder(CVAE) [27, 72], Generative Adversarial Network (GAN) [21, 24], normalizing flow [30, 53], and Transformers [7, 19, 65].

2.1.2 Online prediction. Unlike offline prediction, online prediction renders the output in a sequential manner predicting for each time-step separately. Among the appliance domains of online prediction, the most representative one is the time series forecasting. Time series forecasting has a wide range of applications such as weather forecasting [36, 68], traffic flow forecasting [39, 60], and stock market prediction [33, 62]. Various models based on online prediction can be seen in the literature such as Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), Long Short-term Memory (LSTM), Convolutional Neural Network (CNN), and Temporal Convolutional Network (TCN) [45, 52, 60, 62, 68, 70].

Online prediction can be separated into two types which are sliding window prediction and adaptive online prediction. For sliding window prediction, predictions are made for each time-step in an independent manner with a pre-trained weight without considering its previous output data. Adaptive online prediction also predicts sequentially for every time-step but its predictor's weights are updated for each prediction step. As the prediction of the next step is made based on the previous time stamped data, continuous values are rendered.

For cases where online prediction is applied, such as the time series forecasting, the data is often not available to make the future prediction. To resolve such problem, observations from previous time-steps can be used as input to a regression equation to predict the value at the next time-step. Such technique that predicts by feeding the output back to the model is called to be autoregressive. Both online prediction techniques of sliding window prediction and adaptive online prediction can be autoregressive.

The generation of nonverbal signals is time-dependent like time series problems. As previous SIA behaviors, which are needed to produce its next behavior, are unavailable as in time series forecasting, the aspect of predicting based on the previous time stamped data in an autoregressive manner can be useful for our case. The memory retention present within recurrent networks such as RNN, LSTM and TCN, has shown great promise in time series forecasting. As human behaviors heavily depend on previously performed ones, this aspect of memory is also important for our situation. Moreover, as behavior must be continuous, it is preferable to employ the adaptive online prediction.

2.2 Nonverbal behavior generation models

The generation of the multimodal behavior of SIAs requires to model the temporality of exchanged social signals. Both intrapersonal temporality (coordination of the multimodal communicative behaviors within a single person) and interpersonal temporality (multimodal behaviors arising during dyadic or multi-person interactions) are essential components of the reciprocal adaptation as we adapt our behaviors depending on our prior behaviors and the behaviors shown by others. Previous works that modelize intrapersonal temporality proposed models that generate facial expressions and communicative gestures linked to speech. These works employ Deep Learning (DL) techniques such as Feed-Forward Neural Network (FFN), Bi-directional Long Short-Term Memory (BLSTM), Conditional Variational Autoencoder(CVAE), Generative Adversarial Network (GAN), and Transformers [3, 7, 16, 19, 21, 27, 29, 32, 56, 72]. For our study, we focus on dyadic interactions which leads us to concentrate on modeling the temporal relationship between participants during an interaction. We will look into the literature that considers both interpersonal and intrapersonal temporalities using multimodal signals (only for dyadic interaction).

The modeling of nonverbal behaviors for dyadic interactions started off with rule-based systems such as manually designed rules that were used for predicting backchannels [61], decision trees for chatbot systems generating natural responses and their timing [50], and multimodal probabilistic models that predict backchannels via multimodal signals [46]. The generation of nonverbal behavior such as facial expression, head and body motion started to flourish with the rise of DL models. As far as we are aware, Feng *et al.* [20] were the pioneers to consider the relationship between a human user and a SIA. They generate the agent's facial gestures using the agent's and human's previously predicted facial gestures by creating a Feed-Forward Neural Network (FFN) model. They solely use visual features (facial landmarks) and do not make use of the multimodal information present in the interaction. Also, it is exposed to the problem of outputting discontinuous predictions between two time-steps. Grafsgaard *et al.* [25] learn by encoding the multimodal signals (facial expression, body motion, and speech) using a Long Short-Term Memory (LSTM) model to predict the facial expression and motion of a partner with the speech of both partners and their facial expression and motion features. The interpersonal relationship is modeled by encoding both partners' behaviors; the multimodality is considered but their behavior predictions risk to be not fluid. Dermouche *et al.* [15] also study the interpersonal relationship by referring it as the interactive loop to generate the agent's behavior. They additionally modelize the temporality of nonverbal signals by introducing their Interactive Loop LSTM (IL-LSTM) that considers both agent's and user's upper face behaviors to model the agent's nonverbal behaviors. Similarly, to the model in [20], the IL-LSTM has the same issue of only taking unimodal input features (facial gestures) and as it generates using the sliding window prediction it produce jerky movements. Woo *et al.* [70] address the problem of discontinuous motion prediction of the IL-LSTM in [15] by proposing the use of online LSTM (an adaptive online prediction) which continuously updates memory cells during the whole interaction and leverage multimodal information of visual and acoustic features.

For motion generation, several works use generative models such as Generative Adversarial Network (GAN) [24] and normalizing flow-based models to generate motions that are more diverse and realistic. An extended system of MoGlow [30] is used by Jonell *et al.* [31] to predict the agent's facial expression based on the audio of both partners and the facial expression of the human by encoding all modalities using a RNN and passing their concatenation to a neural network at each time-step of the flow. Tuyen *et al.* [63] forecast the upper body motion (face, body, and hand landmarks) with a context aware model that consists of three components of context encoder, generator, and discriminator. The context encoder encodes the interacting partner's nonverbal behaviors (body motion and audio) and passes the encoded contextual information to the generator along with the target person's body motion. Then the actions outputted by the generator is injected into the discriminator with the contextual information to validate the motion. The two generative models employed in [31, 63] create various possible behaviors by modeling the two facades of interpersonal temporality and multimodality. Nevertheless, they face the same problem of not establishing a continuous link between two sequentially but separately predicted outputs. With the emerging trend of the Transformers model [65], Ng *et al.* [48] generate a continuous 3D facial motion of the listener via an autoregressive transformation-based predictor taking the output of the cross-modal attention that combines the speaker's facial motion and audio inputs and that of Vector Quantised Variational AutoEncoder (VQ-VAE) [64] which discretizes the listener's past facial motion. Their architecture allows the modeling of interpersonal temporality and multimodality, and render continuous predictions via the autoregression. One point that could be hindersome about the model is that transformer-like models require massive amount of data to train. Thus, it might not be suitable for all applications that do not have sufficient amount of data.

The aforementioned models show how the relationship between the interlocutors and the multimodal signals can be modeled. For our work, we want to model the reciprocal adaptation by considering the two facets of temporality (both intrapersonal and interpersonal) and multimodality along with the continuity aspect for the generation of our agent's nonverbal behavior. The multimodality modeling is absent in [15, 20] and the continuity is not assured for [15, 20, 25, 31, 63]. While [48] meets all three of our criteria, it requires a lot of training data. In our case, we have a small database making their model not suitable for our application. We propose a new model structure, in Section 4, that renders continuous nonverbal behaviors (for both speaker and listener) performing with a small dataset. It also learns to capture interpersonal relationship between the interlocutors from the exchanged multimodal signals to endow SIAs with the reciprocal adaptation capability.

2.3 Objective evaluation measures

The evaluation of SIA's non-verbal behavior sequences is a difficult and ill-posed problem: depending on the person, the time of day, our mood, we communicate and react differently to our interlocutor. For example, we may or may not respond to a nonverbal signal (e.g. smile), with more or less intensity and more or less latency. In the same way, head movements are important in maintaining

engagement but they do not obey to strict and precise laws, and a multitude of movements are possible in response to an interlocutor. However, not all occurring movements are perceived as social, convincing, informative or even carrying meaningful information. This is what we want to learn during sequence generation: to generate multimodal behavior sequences that convey the intended intention (e.g. maintaining engagement) and is perceived as such by the human interlocutors.

But how do we evaluate the quality of the generative behaviors models? There is no unanimous answer to this question today. We present here some quantitative measures used in the literature and propose other ones more adapted to our problem. While there is a large literature on subjective measures (see Fitrianie and colleagues' work [22, 23]), we focus on objective evaluation measures in this section.

Behaviors can be interpreted as temporal sequences as their values change in time such as the position in the relative or absolute place for head and body motion or the intensity for Action Units (AUs) [17], which are fundamental actions of facial muscle movements. Thus, we look at sequence comparison measures to evaluate the behaviors in term of accuracy and quality.

One way to assess the accurateness of a generated behavior sequence is to compare its values against the ground truth sequence at each time-step (under the condition that they have the same length). This kind of measure is also often used as a loss function during neural networks learning. Among these measures, we can cite the Mean Squared Error (MSE) [16, 56], the Root Mean Squared Error (RMSE) [15], or the Average Position Error (APE) [1, 2, 29]. Similarly, other authors focus on correlation [16, 25, 56]. These measures allow us to define loss functions for neural network learning by presenting examples of ground truth sequences. However, they cannot be used to evaluate objectively a multimodal sequence generation model for the reasons mentioned above: there is not a fixed behavior (the ground truth) for a given situation and there are multiple plausible answers. In addition, the occurrence timing of a particular behavior can be temporally shifted by a few seconds (for example the behavioral mimicry generally occurs after 2 to 4 seconds [37]) and still be perceived as synchronized [12].

A lot of solutions are used to estimate the quality of sequences generated using Generative Adversarial Network (GAN). They are often based on the principle that several sequences are generated for the same testing example. A solution [31, 42, 56] consists to estimate the distribution over generated sequences and then, to calculate the log-likelihood of the ground truth sequence. Another solution is to measure the smallest distance between the generated sequences and the ground truth one, and average these distances along the testing sequences. Aliakbarian *et al.* [4] estimate the diversity of the generated sequences as the average distance between all pairs of generated sequences. At the same time, they measure the quality using a binary classifier that discriminates between real and generated sequences. Other authors use statistical measures of Inception Score (ID) or Frechet Inception Distance (FID) to measure the generation fidelity of the human motion [5, 11].

All the previous measures assume that several sequences are generated for a same test sequence or that we can estimate the distribution of real sequences. The reciprocal adaptation leads us to a very specific case where the previous measures cannot be applied. More importantly, a lot of temporal dependencies exist between both partners and these phenomena are not observed using the previous measures. Thus, we are also interested in the interpersonal relationship and how to measure it.

While conversing, the speech and movement of the interlocutors are dynamically coordinated (i.e. interpersonal synchrony). However, the detection of such coordination is not so simple as in a real conversation the signals do not happen simultaneously as they result from an exchange. Each interlocutor can send or respond to a signal with a certain time delay (after a perception time [12]). For example, when a person smiles, the interacting person can respond to this smile or not. This response is be perceived as a mimic of the first smile if it happens within a time delay of 2 to 4 seconds [37]. Thus, we need to take into account time shifts. The mimicking behaviors can also differ in terms of duration and intensity. This implies that the sequence comparison also needs to be invariant to dilations when comparing the signals. A well-known technique that deals with such aspects is the Dynamic Time Warping (DTW) [47]. The similarity between two temporal sequences of different speed and length can be measured.

Various efforts have been done to quantify the quality of nonverbal behaviors. Nevertheless, there is not yet a perfect metric to evaluate them. Especially several aspects of behavior quality such as naturalness and human-likeness might be trivial for a human, but still very hard to access for a machine [22, 23]. Thus, human evaluation remains as a critical part of behavior evaluation [3, 11, 13, 20, 23, 31, 32, 56, 72].

3 DATABASE AND FEATURE EXTRACTION

We chose to use the NoXi database [10], which is a corpus of screen-mediated face-to-face interactions containing human-human conversations around a common topic. The database is made up of 3 parts depending on the recording location (France, Germany, and UK). We focus on the recording from the French location that includes 21 dyadic interactions performed by 28 participants with a total duration of 7h22.

We obtain nonverbal behavior features for both interacting participants through feature extraction. For each time-step, the visual features of eye movements (around the x and y axes), head rotations (around the x, y, and z axes), 6 upper face Action Units (AUs) [17] (which are AU1, AU2, AU4, AU5, AU6, and AU7) along with that of the smile (AU12) are extracted via the opensource toolkit OpenFace [6]. The audio features are also obtained for each time-step, after a denoising phase, using the opensource toolkit openSMILE [18]. We consider the following acoustic features: fundamental frequency, loudness, voicing probability, and 13 Mel-frequency cepstral coefficients (MFCCs) [40]. To clean up the data, we apply a median filter and a linear interpolation on all extracted features. All features are adjusted to $25fps$.

4 MODELS

We hold interest in generating social and engaging nonverbal behavior of a SIA (be a speaker or a listener) when interacting with its human interlocutor. In particular, we aim to model the reciprocal adaptation, by capturing the behavior coordination of both interactants, notably the interpersonal relationship. We propose a

new architecture that models the reciprocal adaptation which is our Augmented Self-Attention Pruning (ASAP) model[1], as illustrated in Figure 1. It takes 100 previous frames ($t-99:t$) for both human and agent to predict the agent behavior of the next frame ($t+1$). ASAP consists of three key techniques: data augmentation technique, self-attention pruning, and autoregressive adaptive online prediction.

4.1 Data augmentation

Since our database is not that large, we make use of a data augmentation technique. To learn the reciprocal adaptation we need accurate data of both participants. Which leads us to propose a data augmentation technique which learns from both interlocutors in an equal manner, instead of using classical data augmentation techniques, such as adding noise or dropouts. That is, we learn from the characteristics of both interacting partners. For each batch of the training phase, we assign randomly the interlocutor identity that will be played by the agent to one of the interlocutors. We learn to predict the behaviors for this interlocutor. Then, we follow by alternating and assigning the interlocutor identity for the agent to the other interlocutor and continue the learning process. For a better understanding, we refer to each interacting person of a dyad as *PA* for person A and *PB* for person B. There are two possible choices of giving the agent the interlocutor identity of either *PA* or *PB*. During each batch, the interlocutor identity of the agent is reassigned randomly (to either maintain the same identity of the previous batch or to switch identities from *PA* to *PB* or *PB* to *PA*). The agent learns to generate the behavior of the corresponding interlocutor identity. The data augmentation simulates the interlocutors' behaviors without separating whether it's those of a speaker or a listener. It only takes into account the interlocutor identity (either *PA* or *PB*). By doing so, the model learns to predict equally the behaviors of both participants and focuses on modeling the interaction between the two rather than the specific characteristics of a single person.

4.2 Self-Attention Pruning

To better model the reciprocal adaptation, we want to capture interpersonal relationship (of interpersonal behavior coherence and synchrony) and multimodality from key features. The selection of relevant features is done via an attention mechanism. A self-attention, using the multi-head attention of the Transformers [65], is performed using all the features (2 eyes movements, 3 head rotations, smile (AU12), and 6 upper face AUs) of all interlocutors and (visual and acoustic) modalities. The self-attention layer captures key information to model which behaviors should occur along with mimicry and synchronization mechanisms all at once. However, most attention heads within the multi-head attention (MHA) contain redundant information [44, 66] which lead the model to overfit. Michel *et al.* [44] and Voita *et al.* [66] demonstrate the overfitting problem caused by redundant attention heads can be solved by applying pruning (i.e. pruning removes redundant heads). Our aim is to modelize the reciprocal adaptation, by retrieving key information via pruning. Pruning allows us to drop repetitive heads only rendering attention to dissimilar heads encoded with unique information and it also increases the inference speed. The pruning of attention heads is similar to structured pruning where neurons are pruned. An example of structured pruning is given in Figure 2. Instead of pruning the neurons, we prune the attention heads. Our technique differs from the conventional pruning technique which prunes a given percentage of less significant neurons or connections (for unstructured pruning). Once the model is trained, the same neurons/connections are pruned out disregarding the input. For our pruning technique, we learn to choose which head(s) are meaningful for each specific frame via a pruning mask. For each input sequence, a custom pruning mask is applied. To detail, as seen in Figure 1, the input sequence that consists of $T = 100$ frames from $t-T+1$ to t are passed through the MHA (with the depth of d and N attention heads). A custom pruning mask is learned to minimize the loss of the network for each input sequence to prune the attention heads (each with the dimension of $d \times T$ where d is the depth). The custom pruning mask selects to learn from a certain number of attention heads N' out of N heads. In this way, the pruned attention heads vary for each prediction. For each head, the significance factor is obtained by applying a sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ element-wise and then binarized (by rounding) within the pruning mask. To detail, the pruning mask is a vector of dimension N where each element corresponds to the significance factor of each N MHA heads which is obtained via the sigmoid function. The significance factor is binarized for each element to only leave significant heads as 1 and the rest as 0. Non-significant heads are removed after applying the pruning mask to the attention heads outputted by the MHA. Then, the information of the key heads are grouped together (dimension reduced from $N \times d \times T$ to $N' \times d \times T$) and then the essential information among the information of the key heads are obtained via a fully connected layer (the self-attention pruning module rendering the final dimension of $d \times T$). With our pruning technique, we can assure that our model accesses only unique and relevant information for each prediction.

4.3 Autoregressive adaptive online prediction

We want to generate continuous SIA behaviors which is assured by applying the adaptive online prediction. During the whole course of the interaction the model updates its memory in a continuous way as in [71]. Non-continuous values come from the predictions that are made independently for each input sequence without conserving previous memories (i.e. temporal sliding window). By applying adaptive online prediction during the inference, we circumvent this problem as the past information is kept within the memory cells and used to make new predictions. Also, the prediction is made in an autoregressive fashion by feeding back the predicted values of previous time-steps as input for the prediction at the next time-step.

5 EVALUATIONS

Our goal is to evaluate if ASAP captures the reciprocal adaptation between participants, that is the interpersonal relationship encoded with multimodal signals. Also, we check the quality of our generated SIA behavior with both roles as listener and speaker. We compare the performance of ASAP to that of two recent state-of-the-art models, which are the works of Dermouche *et al.* [15] and Woo *et al.* [70], by evaluating their generated nonverbal behaviors both quantitatively and qualitatively.

[1]The code is available here: https://github.com/jieywoo/ASAP.

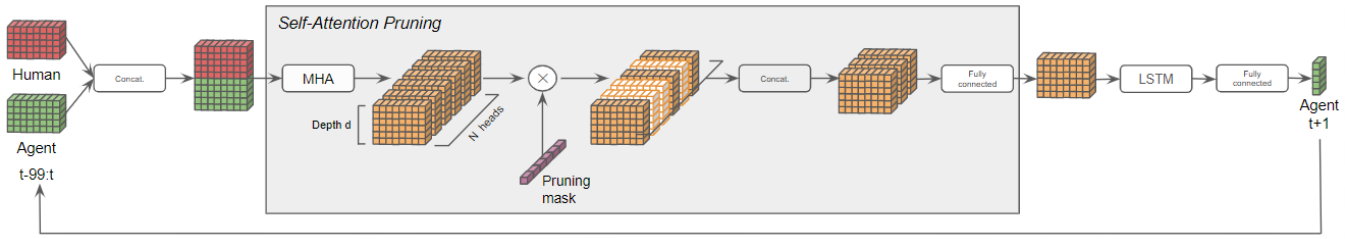

Figure 1: Architecture of ASAP model.

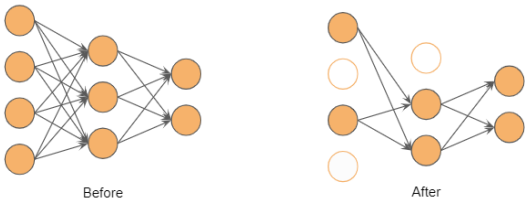

Figure 2: Example of structured pruning.

5.1 Settings

All models are implemented in Tensorflow and trained for 1000 epochs on 2.20GHz Intel Xeon Linux server with NVIDIA GeForce GTX TITAN X and 64GB RAM. They all share the same parameters: batch size of 32 and Adam optimizer with a linear learning rate scheduler (learning rate starting from 0.001, factor 0.2 decay on plateau, and patience 3). For ASAP, after fine-tuning the MHA, four attention heads with the depth of 16 was used and MSE was used as the objective function in the autoregressive stage. The dataset is splitted for training:validation:testing in the ratio of 70:10:20 and ensured that the test set contains pairs of dyads that were never seen in the train and validation sets. To assure that the training and test sets do not include the same person, we have manually excluded participant pairs for the test set.

5.2 Objective evaluation

As mentioned above in Section 2, evaluating nonverbal behaviors has always been a challenge. Until now there is no perfect measure that can thoroughly quantify the dynamics of the behaviors. To assess our model, we propose to use several objective measures, one metric for each measuring type (i.e. point to point, statistical, and resemblance).

5.2.1 Applied measures. As point to point measure, we use the Root Mean Square Error (RMSE), as in the literature [15], to evaluate our generated nonverbal behaviors. The Mean Square Error (MSE) or RMSE between training and testing databases is often used as point to point measure. RMSE, which calculates the error of the generated time series sequence $\hat{y}(t)$ against the real one $y(t)$, is defined by:

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

This measure provides information on the quality of learning. However, it is not always pertinent to compute the exact behaviors that may arise during an interaction, as different reactions (behaviors) of a participant may arise. Indeed, it is difficult to exactly reproduce the same behavior of a person from a database that contains various participants (excluding the targeted person) each possessing a personality and showing different behaviors. We chose to use another measure to further evaluate our model. We are interested to measure if the behaviors generated by our model have similar distributions as in the NoXi database. That is we check if both, predicted behaviors and ground truth have similar number of occurrences. Taking the smile as an example, during the course of a conversation the smile intensity of a participant varies continuously. In the NoXi database, the intensity distribution of smiles is more concentrated around subtle and low level (with the percentage of 84%). We want to assess the quality of the produced nonverbal behaviors globally not on the sequence level but on the entire interaction. Using the example of smiles, we want to see if smiles are predicted through out the interaction in terms of the distribution of smile intensity level. For this purpose, we check the probability distribution similarity using statistical measures.

The quality of nonverbal behaviors can be quantified by verifying their probability distribution. The distribution estimation measures of log-likelihood and density comparison [4, 31, 42, 56] evaluate the difference between predicted and ground truth sequences; but as stated above we want to compare the distribution of the interaction as a whole. In general, when measuring the similarity, statistical measures are used. Aforementioned measures of Inception Score (ID) or the Frechet Inception Distance (FID), which are distribution-based metrics used for scoring the generation fidelity, can not be used for our work since they require an external classifier for ID or to compare the distributions of generated and real objects for FID. As previously presented measures, in Section 2, do not suit our case, we propose the usage of Kolmogorov-Smirnov (KS) two-sample test [43]. Its use is new to behavior quality evaluation. The KS test is a statistical measure that estimates the quality in a quantitative manner by measuring the difference in density probability between the ground truth and the generated sequence for each output dimension. The KS test measures the distance between the generated g(x) and real r(x) data distributions (or more precisely the cumulative distributions G(x) and R(x)):

$$d = \max_x |G(x) - R(x)| \quad (2)$$

The KS test is applied for each feature feature and the average score is calculated.

Point to point metrics and statistical measures for density distribution do not capture the temporal dependencies that exist between partners. To better observe the temporal dependencies between the interlocutors, we employ the Dynamic Time Warping (DTW). DTW measures the similarity between two temporal sequences that may vary in speed and length.

DTW, like the RMSE, can be used between $\tilde{PA}$&PA, where $\tilde{PA}$ is the generated agent's behavior and PA is the human ground truth behavior. Instead of having another precision measure, we want to measure whether the reciprocal adaptation is well captured. The presence of reciprocal adaptation (interpersonal temporal dependency) is verified by seeing if the interlocutors show similar behaviors, responding to each other. We check the proximity (resemblance) of the generated agent's behavior and that of the interacting human ($\tilde{PA}$&PB) and the proximity of the behaviors between both humans (PA&PB) to see if the agent behavior shows the same adaptation trends as seen in the ground truth.

The DTW distance does not have to be small. Actually, it would be easy to copy the behavior of the human at the previous moment to have a DTW almost zero. This high resemblance between partners can be perceived as an everlasting imitation (like a parrot) and thus may rather hinder the perception of human-like behavior. Thus, DTW between $\tilde{PA}$&PB must be similar to the DTW between PA&PB and not necessary small.

As our interactions are very long (around $20min$ for each interaction), we compute the DTW in small chunks of $1min$ and a stride of $30s$. Applying DTW in chunks speeds up the computation. All the chunks cover the whole interaction.

Smile is an key socio-emotional signal that can be observed frequently during an interaction [35]. Previous studies have demonstrated that smile helps SIAs to better manage their interaction with their human users [51, 69]. Thus, for DTW distance evaluation of $\tilde{PA}$&PB, we focus on the smile.

5.2.2 Results and discussions. To compare our model with that of the literature, we need to use the same features. As a result, we firstly evaluate our model with the features presented in [15] (features set 1) and then with those in [70] (features set 2). The features set are composed as the following:

- Features set 1: only visual features (eyes movement, head rotation, and AU12 intensity and activation) of both interlocutors along with conversational state inputted to predict visual features of the SIA at $5fps$;
- Features set 2: visual and acoustic features (eyes movement, head rotation, upper face AUs and AU12 intensities, fundamental frequency, loudness, voicing probability, and 13 MFCCs) of both interlocutors to predict the visual features (including upper face AUs) of the SIA at $25fps$.

Concerning the evaluation of the eyes movement, we evaluate the value of the eyes angles like we do for the head rotation. However, we cannot assess if the predicted eyes movement correspond to looking at the same target (e.g. its interlocutor) as in the ground truth as this information is not available in the NoXi dataset (both cameras recording the two interlocutors are not calibrated).

Methods		RMSE	KS test
Features set 1	Dermouche *et al.*, 2019	0.172	0.298
	Woo *et al.*, 2021	0.171	0.293
	ASAP	**0.131**	**0.115**
Features set 2	Dermouche *et al.*, 2019	0.444	0.559
	Woo *et al.*, 2021	0.374	0.415
	ASAP	**0.239**	**0.301**

Table 1: Average RMSE and KS test results for features set 1 and 2.

All models were trained and their behaviors were generated for each features sets. We conduct an objective evaluation for the two sets of features.

The performance of ASAP is compared with the baseline models for each features set using the proposed objective evaluation measures. In Table 1, the three models of each features set are evaluated quantitatively by computing the RMSE and performing the KS two-sample test. The KS test was used as it statistically measures the probability distribution similarity between our predictions and ground truth (real interaction). The average score of the output features is calculated (average of 6 output features scores (2 eyes angles, 3 head rotations, and AU12 intensity) for features set 1 and that of 12 output features scores (2 eyes angles, 3 head rotations, and the intensities of 6 upper face AUs and AU12) for features set 2). From both features set 1 and 2, we can observe that the RMSE and the KS test scores have better values for ASAP than the baseline models.

The DTW between PA&PB represents distance (resemblance) between the signals of the two human participants' interlocutor identities of PA and PB. The DTW distance is interpreted as the closer the distance gets, the more the two signals of PA and PB are similar. We check if the models' DTW distance $\tilde{PA}$&PB is close to that of the ground truth interaction (human-human interaction) PA&PB.

As stated above, smile is a key social signal that is apparent to improve SIA's interaction which leads us to focus on smile. We can see, in Table 2, that for smile of features set 1, our ASAP performs better than the baseline models in terms of having the DTW distance the closest to the ground truth DTW (26.9, 21.7 respectively). The same conclusion can be drawn for features set 2 (1399.3, 1317.5 respectively). Note that the small value of obtained with Woo *et al.* model can be interpreted as a close imitation of the behavior of its interlocutor that may deter the perception of the behavior to be human-like.

Therefore, we can conclude that our ASAP model outperforms the baseline models for the three objective evaluation methods, that is RMSE, KS test, and resemblance via DTW distance between PA&PB.

5.3 Subjective evaluation

Relying only on objective evaluations is not enough to fully assess the quality of the generated agent's behavior. We perform a user perceptive study to complement the objective evaluation where we look more particularly on how the generated multimodal signals and the modeling of the reciprocal adaptation (interpersonal relationship) by our model influence: 1) the perception of the generated agent

Features set	Method	DTW *PA&PB* (Ground truth)	DTW $\tilde{PA\&PB}$
Features set 1	Dermouche *et al.*, 2019	21.7	27.3
	Woo *et al.*, 2021		27.2
	ASAP		**26.9**
Features set 2	Dermouche *et al.*, 2019	1317.5	1562.7
	Woo *et al.*, 2021		257.5
	ASAP		**1399.3**

Table 2: DTW of smile for features set 1 and 2.

behaviors' naturalness and human-likeliness; 2) the perception of the interpersonal dynamics such as the synchrony between the interlocutors and the perception of their engagement. To evaluate these aspects of human-agent interaction, we ask the participants to score the interacting SIA along 4 measurement constructs: behavior naturalness, behavior human-likeliness, interaction synchrony, and engagement.

Questionnaires to evaluate the perception of behavior naturalness (e.g. "Is the behavior of the virtual character artificial?"), behavior human-likeliness ("Does the virtual character behave like a human?"), and engagement ("Is the virtual character engaged in the conversation?") are formulated based on existing questionnaires of human-agent interaction evaluation [23, 67]. We use a set of three synonyms and antonyms for each dimension. To evaluate the perception of synchrony, we use the dyadic stances of mutual understanding, attention, agreement, interest, and pleasantness (e.g. "Are the human and the virtual character agreeing to each other?") proposed by [41, 55].

A set of 14 questions (3 for each construct of behavior naturalness, behavior human-likeliness, and engagement, and 5 for interaction synchrony) are used. The users are asked to answer to each question using a Likert scale of 5 points (ranging from 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), to 5 (strongly agree)).

5.3.1 Hypothesis. In human-human interaction, behaviors are interpersonally coordinated [57], which is also referred to as interpersonal synchrony [14]. It is also shown that being in sync improves the engagement level [14, 28]. Thus, we hypothesize that our ASAP model improves interpersonal dynamics (synchrony and engagement) of the generated agent behaviors as well as its quality (naturalness and human-likeliness) compared to the baseline models. The perceptive study enables us to validate the hypothesis that our model generates a more interactive and adaptive SIA for dyadic conversations.

5.3.2 Procedure. The evaluation is done via Prolific, an online crowd-sourcing platform. 20 video clips of an approximate duration of 7 seconds are extracted from the human-human videos of NoXi. In each video clip a human participant has the speaking turn (talking about a common subject) or is the listener (expressing nonverbal behaviors with visual and acoustic feedbacks which include backchannels such as "ok" and "yes") and the other human participant is either, respectively, listener or speaker.

For our study, we compare four conditions (the three models which are: ASAP and our two baseline models of Dermouche *et al.* and Woo *et al.*) with the features set 2 and the ground truth). To evaluate the quality of these conditions, we replace one of the human participant (being the speaker in 10 video clips and listener in the other 10 video clips) by a SIA whose behavior is driven by the computational models or the ground truth. The SIA was animated using the open source Greta SIA platform [49] by passing visual features (predictions of the computational models or the ground truth) along with the audio of the ground truth. An image of a video is shown in Figure 3 in which it displays a SIA (left side of the screen) and a human participant (right side of the screen). The lower face was blurred so that the mouth movements won't hinder the evaluators' perception during the study.

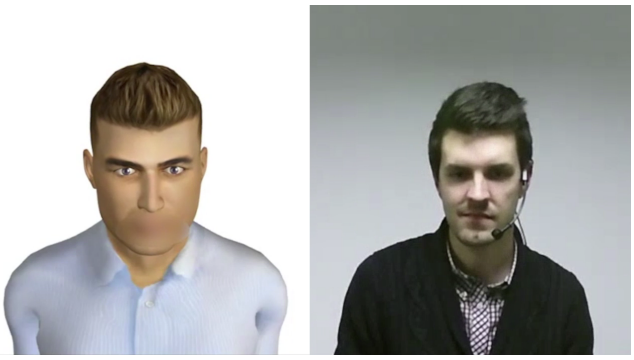

Figure 3: User perception test video clip example of an interaction between a SIA (left) and a human participant (right).

Four videos (the agent displaying the behavior of the agent in one of the four conditions) are created for each of the 20 human-human video clips of the NoXi database. So, we have a total of 80 videos where the SIA replaces one of the human interlocutors (see Figure 3). The behaviors of ground truth condition is also shown by replacing the selected human with the SIA. We use the same setting when comparing videos of the ground truth with videos of the computational models. As such we eliminate any impact a participant may have toward the virtual character [58].

Not to make an evaluation that lasts too long which may deteriorate the concentration of the perception study participants and thus hinder the study, we split the perception test into four groups. Each group has 5 human-human interaction video clips to evaluate (i.e. each participant evaluates 20 short videos of 7s of human-agent interaction for all four conditions). All the videos are shuffled so that their order does not impact our perception study.

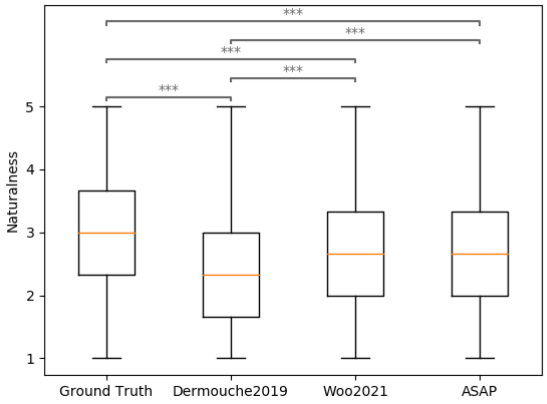

Figure 4: Distribution of behavior naturalness ($^{***}p < 0.001$).

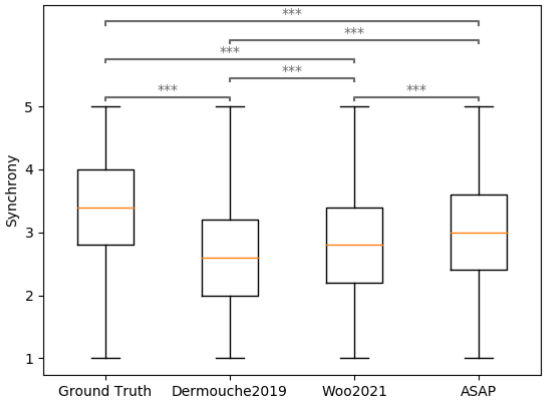

Figure 6: Distribution of synchrony ($^{***}p < 0.001$).

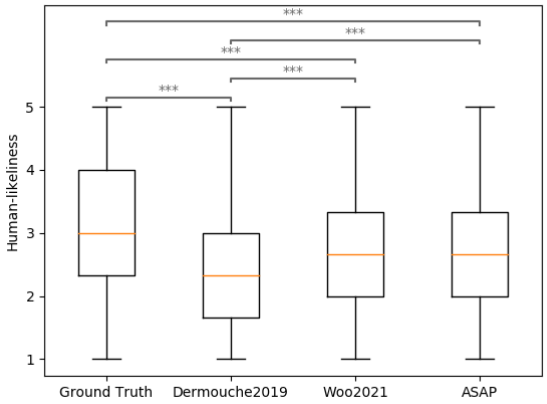

Figure 5: Distribution of behavior human-likeliness ($^{***}p < 0.001$).

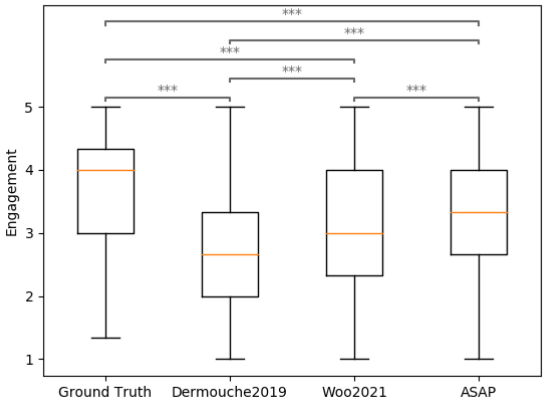

Figure 7: Distribution of engagement ($^{***}p < 0.001$).

For each perception test group, we recruit 30 participants and ask them to evaluate each video (20 videos per group) with the aforementioned set of questions. To filter out inattentive participants, for each video we randomly include attention check questions (e.g. "Is the virtual character playing tennis with the human interlocutor?").

5.3.3 Results and discussions. The participants' responses are grouped together according to their corresponding construct (behavior naturalness, behavior human-likeliness, synchrony, and engagement) for each condition (ground truth, our two baseline models of Dermouche *et al.* and Woo *et al.*, and ASAP). We visualize the distribution for each construct, in Figure 4, 5, 6, and 7.

One-way ANOVA report significant differences among all animation conditions for all four constructs: behavior naturalness ($F = 41.5, p < 0.001$), behavior human-likeliness ($F = 43.1, p < 0.001$), synchrony ($F = 66.9, p < 0.001$), and engagement ($F = 90.0, p < 0.001$). A post-hoc pairwise comparison analysis is performed by running the Tukey's honestly significantly differenced (HSD) test. Tukey's HSD reveal the following. Statistical significant differences were found between all pairs ($p < 0.001$) except between Woo *et al.* [70] and ASAP for the constructs of behavior naturalness and human-likeliness ($p = 0.9$ and $p = 0.9$ respectively). Concerning the constructs of synchrony and engagement, all pairs were reported to be significantly different ($p < 0.003$). A two-tailed t-test was performed between all possible pairs of compared animations for each construct to test the statistical significance. The t-test p-values reported significant differences between all pairs ($p < 0.001$) except between Woo *et al.* [70] and ASAP for the constructs of behavior naturalness and human-likeliness ($p = 0.7$ and $p = 0.5$ respectively). T-test yields significant differences among all conditions for synchrony and engagement constructs. The t-test p-values are shown on the construct distribution figures (Figures 4, 5, 6, and 7).

From the subjective results, the simulation with ground truth values receives the highest values for all four constructs, namely behavior naturalness (3.0), behavior human-likeliness (3.0), synchrony (3.4),

and engagement (4.0). Via the constructs of behavior naturalness and human-likeliness, a rise in quality can be noticed between that of Dermouche *et al.* (2.3, 2.3 respectively) and the other two computational models of Woo *et al.* (2.7, 2.7 respectively) and our ASAP model (2.7, 2.7 respectively). We assume that this difference is due to the application of adaptive online prediction, instead of sliding window prediction as in Dermouche *et al.*, which enables the generation of continuous motions which may lead to a higher perception of naturalness and human-likeliness. The quality of the generated agent behavior along the constructs of synchrony and engagement increases from the Dermouche *et al.* (2.6, 2.7 respectively), to Woo *et al.* (2.8, 3.0 respectively), to ASAP (3.0, 3.3 respectively). We can remark that modeling of reciprocal adaptation allows SIA to be more in sync and engaged with it's interlocutor.

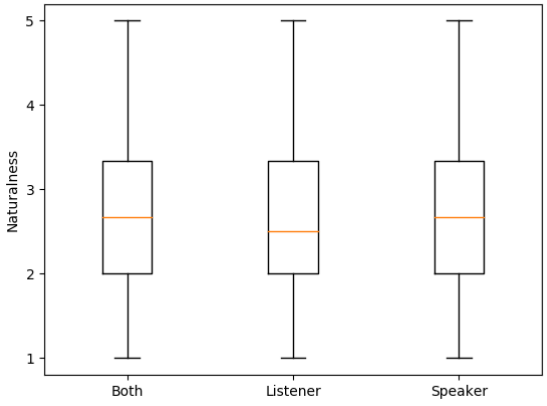

Figure 8: Distribution of behavior naturalness.

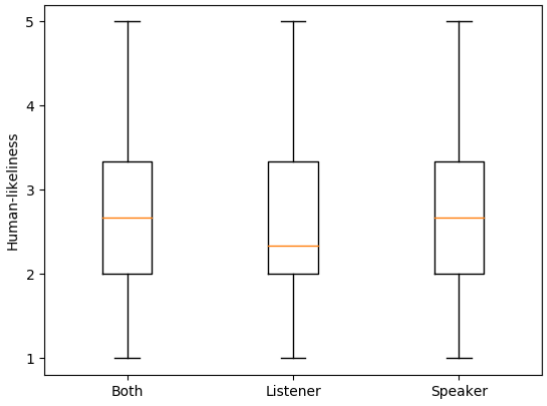

Figure 9: Distribution of behavior human-likeliness.

We also want to evaluate if our ASAP model can produce behaviors for SIA being both a listener and a speaker. We check the quality of the generated agent behavior of ASAP along the four constructs

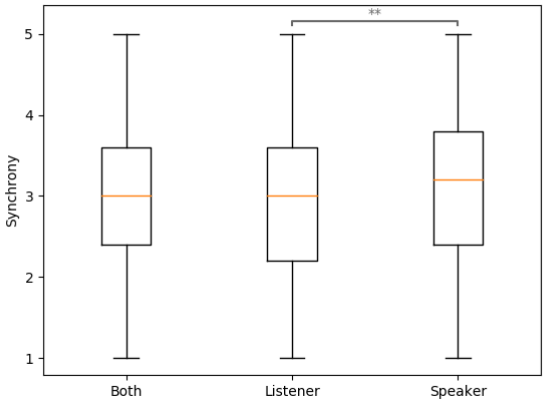

Figure 10: Distribution of synchrony ($^{**}p < 0.01$).

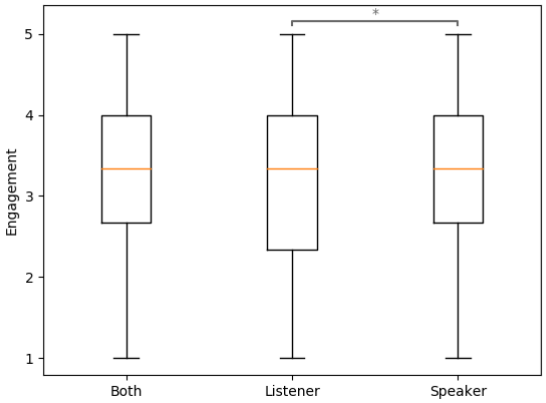

Figure 11: Distribution of engagement ($^{*}p < 0.05$).

by comparing the produced behaviors as a listener and a speaker, as shown in Figure 8, 9, 10, and 11.

For the SIA being either a listener or a speaker or both combined, one-way ANOVA reported significant differences for the construct of synchrony ($p = 0.02$) but no significance for the other three constructs of behavior naturalness, behavior human-likeliness, and engagement. Tukey's HSD on synchrony revealed significant difference between listener and speaker ($p = 0.01$). A two-tailed t-test was performed and showed significant differences between listener and speaker for the constructs of synchrony ($p = 0.005$) and engagement ($p = 0.02$) as indicated on the construct distribution figures (Figures 10 and 11).

We can remark that ASAP generates both listener (2.5, 2.3, 3.0, 3.3 respectively) and speaker (2.7, 2.7, 3.2, 3.3 respectively) behaviors with similar qualities which indicates that ASAP can be used to generate SIA behaviors for an entire interaction.

Our subjective evaluation results are inline with the results of the objective evaluation. Our ASAP model performs better than that of the baseline models of Dermouche *et al.* and Woo *et al.*. Thus, ASAP outmatches the baselines, notably in terms of synchrony and engagement, and is the most similar to the ground truth both quantitatively and qualitatively. Moreover, ASAP can serve to produce SIA behavior for both speaker and listener.

6 CONCLUSION

Having the goal to create an expressive SIA capable of interacting with the user while maintaining his/her attention, we develop a predictive model that produces the agent's nonverbal behaviors serving as both active speaker and listener. We modelize the reciprocal adaptation of our ASAP model by focusing on the aspects of interpersonal temporality, multimodality by encoding multimodal signals, and behavior prediction continuity with the autoregressive adaptive online prediction. Our model outperforms the baseline models through both objective and subjective evaluations. ASAP shows great promise in rendering natural and human-like behaviors that are engaging and in sync with the interlocutor. As for our next step, we aim to better modelize the reciprocal adaptation between the two interlocutors by modeling each interacting partner's intrapersonal temporality along the multimodality aspect and also capturing their interpersonal temporality to generate behaviors that are livelier. In a near future, we intend to assess the performance of the SIA with reciprocal adaptation capability through live human-agent interaction. Another direction of research is to study the aspect of explicability of the reciprocal adaptation to see how the dynamics of the adaptation play in the behavior generation as the conversation evolves.

7 PRACTICAL AND SOCIAL IMPLICATIONS

In this paper, we discussed the endowment of reciprocal adaptation capability to SIAs. The endowment of such capacity has shown an increase in the perception of the SIA's naturalness, human-likeliness, conversational engagement and synchrony. In a more practical point of view, such SIAs may provide better and a wider range of services. By showing adaptation skills, SIAs adapting to its interaction user can enhance the perception of interaction quality, the relation created with its interlocutor, etc. This renders a positive impact on the SIA's task performance (e.g. for a tutoring system, the learning is reinforced [34]). SIAs can be used for a variety of applications and improve people's lives by providing assistance and support. However, with the development of intelligent UIs, some people may tend to avoid human-human interaction and rely on them. This shrink of interactions with other humans is not healthy and is not what SIAs were designed for. To avoid such happening, we should always have in mind that SIAs are here to assist us and not to replace human interaction.

ACKNOWLEDGMENTS

This work is performed as a part of ANR-JST-CREST TAPAS (ANR-19-JSTS-0001) and IA ANR-DFG-JST Panorama (ANR-20-IADJ-0008) 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 Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

[4] Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. 2021. Contextually plausible and diverse 3d human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11333–11342.

[5] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. 2020. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5223–5232.

[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] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents** This work has been supported in part by ARO Grants W911NF1910069 and W911NF1910315, and Intel. Code and additional materials available at: https://gamma. umd. edu/t2g. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 1–10.

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

[9] Judee K Burgoon, Lesa A Stern, and Leesa Dillman. 1995. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press.

[10] 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

[11] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. 2021. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11645–11655.

[12] Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology* 76, 6 (1999), 893.

[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] 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.

[15] 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.

[16] Chuang Ding, Lei Xie, and Pengcheng Zhu. 2015. Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications* 74, 22 (2015), 9871–9888.

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

[18] 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.

[19] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2022. Transformer Network for Semantically-Aware and Speech-Driven Upper-Face Generation. In *EUSIPCO*.

[20] 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.

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

[22] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, and Willem-Paul Brinkman. 2021. Questionnaire Items for Evaluating Artificial Social Agents-Expert Generated, Content Validated and Reliability Analysed. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 84–86.

[23] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 unifying questionnaire constructs of artificial social agents: An iva community analysis. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[24] 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.

[25] Joseph Grafsgaard, Nicholas Duran, Ashley Randall, Chun Tao, and Sidney D'Mello. 2018. Generative multimodal models of nonverbal synchrony in close relationships. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 195–202.

[26] 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.

[27] David Greenwood, Stephen Laycock, and Iain Matthews. 2017. Predicting head pose from speech with a conditional variational autoencoder. ISCA.

[28] Aman Gupta, Finn L Strivens, Benjamin Tag, Kai Kunze, and Jamie A Ward. 2019. Blink as you sync: Uncovering eye and nod synchrony in conversation using wearable sensing. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 66–71.

[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] 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.

[32] 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.

[33] Kyoung-jae Kim. 2003. Financial time series forecasting using support vector machines. *Neurocomputing* 55, 1-2 (2003), 307–319.

[34] Yanghee Kim, Jeffrey Thayne, and Quan Wei. 2017. An embodied agent helps anxious students in mathematics learning. *Educational Technology Research and Development* 65, 1 (2017), 219–235.

[35] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal communication in human interaction*. Cengage Learning.

[36] Neeraj Kumar and Govind Kumar Jha. 2013. A time series ann approach for weather forecasting. *Int J Control Theory Comput Model (IJCTCM)* 3, 1 (2013), 19–25.

[37] N Pontus Leander, Tanya L Chartrand, and John A Bargh. 2012. You give me the chills: Embodied reactions to inappropriate amounts of behavioral mimicry. *Psychological science* 23, 7 (2012), 772–779.

[38] Bo Li, Tara N Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yanghui Wu, and Kanishka Rao. 2018. Multi-dialect speech recognition with a single sequence-to-sequence model. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4749–4753.

[39] Marco Lippi, Matteo Bertini, and Paolo Frasconi. 2013. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems* 14, 2 (2013), 871–882.

[40] Beth Logan. 2000. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*. Citeseer.

[41] Max M Louwerse, Rick Dale, Ellen G Bard, and Patrick Jeuniaux. 2012. Behavior matching in multimodal communication is synchronized. *Cognitive science* 36, 8 (2012), 1404–1426.

[42] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. 2021. Generating Smooth Pose Sequences for Diverse Human Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13309–13318.

[43] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.

[44] Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems* 32 (2019).

[45] Mohsen Mohammadi, Faraz Talebpour, Esmaeil Safaee, Noradin Ghadimi, and Oveis Abedinia. 2018. Small-scale building load forecast based on hybrid forecast engine. *Neural Processing Letters* 48, 1 (2018), 329–351.

[46] 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.

[47] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.

[48] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning to Listen: Modeling Non-Deterministic Dyadic Facial Motion. *arXiv preprint arXiv:2204.08451* (2022).

[49] Radoslaw Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. 2009. Greta: an interactive expressive ECA system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 1399–1400.

[50] 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

[51] Magalie Ochs and Catherine Pelachaud. 2013. Socially aware virtual characters: The social signal of smiles. *IEEE Signal Processing Magazine* 30, 2 (2013), 128–132.

[52] Alfonso Palmer, Juan Jose Montano, and Albert Sesé. 2006. Designing an artificial neural network for forecasting tourism time series. *Tourism management* 27, 5 (2006), 781–790.

[53] 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).

[54] Helmut Prendinger and Mitsuru Ishizuka. 2005. THE EMPATHIC COMPANION: A CHARACTER-BASED INTERFACE THAT ADDRESSES USERS'AFFECTIVE STATES. *Applied artificial intelligence* 19, 3-4 (2005), 267–285.

[55] Ken Prepin, Magalie Ochs, and Catherine Pelachaud. 2013. Beyond backchannels: co-construction of dyadic stancce by reciprocal reinforcement of smiles between virtual agents.. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35.

[56] 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.

[57] Richard C Schmidt and Michael J Richardson. 2008. Dynamics of interpersonal coordination. In *Coordination: Neural, behavioral and social dynamics*. Springer, 281–308.

[58] Youssef Shiban, Iris Schelhorn, Verena Jobst, Alexander Hörnlein, Frank Puppe, Paul Pauli, and Andreas Mühlberger. 2015. The appearance effect: Influences of virtual agent features on performance and motivation. *Computers in Human Behavior* 49 (2015), 5–11.

[59] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27 (2014).

[60] Yongxue Tian and Li Pan. 2015. Predicting short-term traffic flow by long short-term memory recurrent neural network. In *2015 IEEE international conference on smart city/SocialCom/SustainCom (SmartCity)*. IEEE, 153–158.

[61] 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.

[62] Avraam Tsantekidis, Nikolaos Passalis, Anastasios Tefas, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. 2017. Forecasting stock prices from the limit order book using convolutional neural networks. In *2017 IEEE 19th conference on business informatics (CBI)*, Vol. 1. IEEE, 7–12.

[63] Nguyen Tan Viet Tuyen and Oya Celiktutan. 2022. Context-Aware Human Behaviour Forecasting in Dyadic Interactions. In *Understanding Social Behavior in Dyadic and Small Group Interactions*. PMLR, 88–106.

[64] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[66] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418* (2019).

[67] Astrid M Von der Pütten, Nicole C Krämer, Jonathan Gratch, and Sin-Hwa Kang. 2010. "It doesn't matter what you are!" explaining social effects of agents and avatars. *Computers in Human Behavior* (2010).

[68] Renzhuo Wan, Shuping Mei, Jun Wang, Min Liu, and Fan Yang. 2019. Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting. *Electronics* 8, 8 (2019), 876.

[69] Isaac Wang and Jaime Ruiz. 2021. Examining the use of nonverbal communication in virtual agents. *International Journal of Human–Computer Interaction* 37, 17 (2021), 1648–1673.

[70] Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. 2021. Creating an interactive human/agent loop using multimodal recurrent neural networks. In *WACAI 2021*.

[71] 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

[72] Ye Yuan and Kris Kitani. 2020. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*. Springer, 346–364.