Pattern Recognition Letters Learning Off-line vs. On-line Models of Interactive Multimodal Behaviors with Recurrent Neural Networks

Duc Canh Nguyen, Gérard Bailly, Frédéric Elisei

Pattern Recognition Letters
journal homepage: www.elsevier.com

Learning Off-line vs. On-line Models of Interactive Multimodal Behaviors with Recurrent Neural Networks

Duc-Canh Nguyen[a] ∗, Gérard Bailly[a,] and Frédéric Elisei[a]

[a]*GIPSA-Lab, Univ. Grenoble-Alpes/CNRS, Speech & Cognition Department, France*

ABSTRACT

Human interactions are driven by multi-level perception-action loops. Interactive behavioral models are typically built using rule-based methods or statistical approaches such as Hidden Markov Model (HMM), Dynamic Bayesian Network (DBN), etc. In this paper, we present the multimodal interactive data and our behavioral model based on recurrent neural networks, namely Long-Short Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models. Speech, gaze and gestures of two subjects involved in a collaborative task are here jointly modeled. The results show that the proposed deep neural networks are more effective than the conventional statistical methods in generating appropriate overt actions for both on-line and off-line prediction tasks.

Keywords: Face-to-face interaction, multimodal behavior, co-verbal behavior, behavioral models, multi-task RNN, LSTM, Bi-directional LSTM

∗ Corresponding author. Tel.: +33-758-692-625; e-mail: duc-canh.nguyen@gipsa-lab.fr

1. Introduction

Face-to-face communication is one of the most natural and effective form of human communication in our daily life. Modeling human-to-human multimodal interactive behavior is one of the prerequisites to endow artificial agents – virtual avatars or social robots – with conversational skills. There are classically two main approaches to this challenging issue: rule-based vs. machine learning methods.

In data-driven rule-based methods, researchers analyze the recordings of human interactions and try to semi-automatically find patterns in multimodal streams. Computational frameworks are then proposed to operationalize those findings. Such systems usually incorporate set of rules that map perceptual cues to multimodal actions via an intermediate estimation of communicative intentions. For example, the BEAT system [1] generates nonverbal behaviors from text by enriching the linguistic structure with language tags such as rheme/theme contrasts, objects and actions. Lee and Marsella [2] similarly propose a Nonverbal Behavior Generator system to generate behaviors according to communicative functions. Thorisson [3] further proposes an event-based language where a *finite state machine* (FSM) describes an interaction scenario as a series of states with pre-conditions and post-actions structured in three hierarchical layers (reactive, process and content). However, hand-crafted rules have difficulty in taking into account the many factors conditioning the multimodal behaviors (task, personality, social context, emotion, gender, etc.) while maintaining a fine-grained life-like variability.

Another popular approach is based on machine learning techniques which try to find behavior regularities directly from data. For example, Otsuka et al. [4] proposed Dynamic Bayesian Networks (DBN) to estimate addressing and turn taking ("who responds to whom and when?") while using the conversational regime as a latent variable. Mihoub et al [5] estimated interaction units using Hidden Markov Models (HMM) to generate the gaze of an interlocutor from his own speech activity together with the gaze and speech activity of his partner. Mihoub et al [6] further showed that DBN outperform both full- and semi-HMM in predicting co-verbal behaviors in a "put that there" game. Actually, few works have been devoted to the modeling of joint behaviors while incredible amount of research have been successfully dealing with recognition of human activities from multimodal behaviors [7] [8] and vice-versa [9] [10].

2. State of the art: predicting interactive behaviors with RNN

Recently, Recurrent Neural Networks (RNN) have been shown to outperform statistical models in sequence recognition. Gated recurrent units (GRUs) and Long-Short Term Memory (LSTM) cells have been introduced to cope with long-term temporal dependencies. These cells add gates to inputs – and outputs for LSTM – of the processing units. These overcome the vanishing problem of simple RNN. Because of their ability to modulate between short- and long-term dependencies, they are particularly suited for building latent spaces that mediates input-to-output co-variations. Therefore, LSTM becomes state of art of many applications related to sequential data such as statistical language modeling [11], machine translation [12], and description generation from image [13], etc. Another advantage of LSTM is that it can learn timing intervals between sub-patterns in sequences [14]. Such coordination patterns are particularly crucial to multimodal behaviors such as those involved in natural human robot interaction (see section 5.2).

Most of LSTM-based models have been proposed so far for the recognition of human activities. For example, Ordóñez et al combined Convolution Neural Network (CNN) with LSTM to build a DeepConvLSTM framework which is able to recognize human activities from wearable sensors with minimal pre-preprocessing [15]. Furthermore, Tsironi et al also build a CNN-LSTM to learn gestures which have varying duration and complexity [16]. Tian et al [17] performed successful emotional recognition in spontaneous dialogue with LSTM.

Fewer works have been devoted to the generation of interactive behaviors. Alahi et al [18] used LSTM with social pooling of hidden states which combines the information from all neighboring states to predict human trajectories in crowded space. Ravichandar et al [19] built a promising model of sequential tasks using LSTM in order for one robot to predict what human will do next. LSTM-based conversation models [Joty, S. and Hoque, E., 2016] have also recently proposed to predict turns in two-party conversations.

In this paper, we present multimodal interactive behavioral models based on recurrent neural networks, namely Long-Short Term Memory (LSTM) RNN and Bidirectional LSTM (BiLSTM), that predict gaze and arm gestures in a collaborative human-human task.

Figure 1: An example of the "put that there" interaction filmed by a camera placed on instructor's head [6]

3. Interactive data

The dataset[1] used as interactive data in this paper has been collected by Mihoub et al [6]. This face-to-face interaction involves an instructor and a manipulator who performed a collaborative task called "put that there". The experimental setting is shown in Figure 1. In this scenario, the manipulator should move cubes from a reservoir to a chessboard, following instructions given by the instructor. The instructor is the only one to know the target arrangement of the cubes, while the manipulator is the only one being able to move the cubes. Therefore, this task requires the instructor and manipulator to cooperate: share knowledge and coordinate their sensory-motor abilities Each of our instructor/manipulator dyads performed 10 games consisting in reproducing a target arrangement of ten cubes, with an implicit control of the gaze and hand gestures thanks to the initial and final disposition of the cubes. This balanced statistical coverage of behaviors provides an interesting benchmark to collect human strategies used to maintain mutual

[1] http://www.gipsa-lab.fr/projet/SOMBRERO/data.html

attention and coordinate multimodal deixis (finger pointing, head, gaze, etc.) towards objects and locations.

The data here include 30 games performed by one instructor and 3 different partners, in order to replicate an arrangement of 10 cubes on the chessboard from an initial random layout in the reservoir. The mean duration of a game is around 80 seconds. The total duration of interactive data is about 30 minutes.

The interactive audiovisual data were complemented by motion capture (gestures as well as eye tracking). All raw streams are resampled at 25Hz. Additional annotations were performed using Elan [20] and Praat [21]. The final observations consist of 5 streams of discrete variables:

- IU: each game is further segmented into interaction units – that could be also termed as elementary skills or sub-tasks – describing the sequential organization of a repetitive elementary interaction. We distinguish between 6 different IUs mirroring the activities of the instructor: get instruction from tablet, seek the cube to be manipulated, point the cube, indicate target position of the cube, check the manipulation and validate the result. These IU pace the activities of both agents that are characterized by the following observations:
- MP: Manipulator gestures with: rest, grasp, manipulate, put, none
- SP: speech of instructor about: manipulated cube, reference cube, relative positioning, else, none
- GT: region of interest pointed by the instructor's index: rest, manipulated cube, target location, reference cube or none
- FX: region of interest fixated by the instructor's gaze: manipulator's face, reservoir, task space, manipulated cube, target location, reference cube, tablet, else.

The challenge is to predict the instructor's co-verbal gestures GT and FX given his verbal activity (SP) and the interlocutor's gestures (MP). The behavioral models proposed below should thus generate endogenous co-verbal behaviors from endogenous verbal behaviors and exogenous percepts.

4. Training regression models

We compare below the performance of three regression models in predicting endogenous co-verbal behaviors: Discrete Hidden Markov Model (DHMM), Dynamic Bayesian network (DBN) vs. Long-short term memory (LSTM) recurrent neural network (RNN). We tested two versions of each model: (1) off-line models that perform estimations once the whole sequence has been observed and (2) on-line models that perform estimations incrementally at each time frame.

4.1. Hidden Markov Models

A multimodal interactive model based on HMM was proposed in [5]. In this model, each interactive unit (IU) is modeled by one Discrete Hidden Markov Model (DHMM) that models joint multimodal sensorimotor behaviors via its hidden states. Eq. (1) defines parameters of the DHMM models

$$\lambda_p = (A_p, B_p, \pi_p) \quad (1)$$

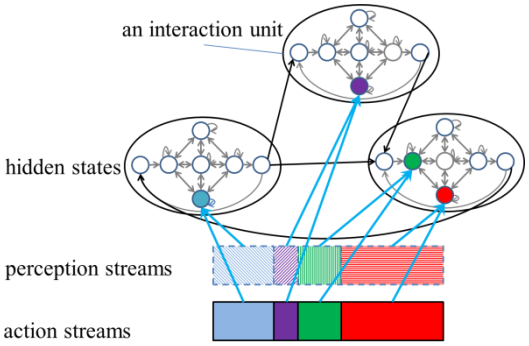
where $p = 1 \ldots P$ is the index of the interaction unit (here P = 6 corresponding to 6 IUs)

The observation vectors (with T is length of the observation sequence) are separated in two parts: the perceptual streams and the action streams illustrated in eq. (2)

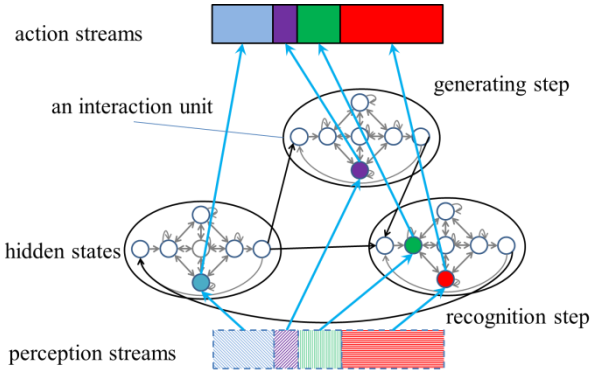
$$O^p = (o_t^p)_{t=1..T} = (SP_t, MP_t)_{t=1..T} \quad (2.a)$$

$$O^a = (o_t^a)_{t=1..T} = (GT_t, FX_t)_{t=1..T} \quad (2.b)$$

$$O = (o_t)_{t=1..T} = (o_t^p, o_t^a)_{t=1..T} = (SP_t, MP_t, GT_t, FX_t)_{t=1..T} \quad (2.c)$$



(a) Training with joint perception and action streams



(b) Estimating hidden states from perception stream and then generating action stream

Figure 2. Schematic of HMM-based multimodal interactive modeling: (a) training (b) generating. Each interaction unit is modeled by a DHMM with fully-connected states. The syntactic organization of these elementary interaction units is described by a fully-connected bi-gram model. Transition probabilities within DHMM and between DHMMs are drawn with gray and black arrowed lines respectively. Cyan arrows represent emission probabilities connecting states with perception and action streams. They are simultaneously trained at training stage (a). At generation (b), perception streams are used to estimate distributions of hidden states and action streams.

Each DHMM is trained using Expectation and Maximization (EM) algorithm. The DHMMs are trained with streams aligned with IUs. Transition probabilities between the DHMMs – i.e. between their input and output states – are described by a bi-gram model, i.e. a fully connected transition matrix, notably because repetitions of IUs or of couples of IUs are sometimes necessary to fulfill the task. At training stage, all data streams are available, while in testing only the endogenous verbal stream and exogenous observations are available as shown in Figure 2a.

After training, two sub-models (a hidden state decoder and an action generator) are thus extracted and used in two steps as shown in Figure 2b. Firstly, the hidden state decoder estimates sensorimotor states from perceptual observations only shown in Eq.(3). The decoding of sensorimotor state sequence is performed offline by Viterbi alignment and online by a bounded Short-Time Viterbi algorithm with no look-ahead.

$$S^* = \arg\max_S P(S \mid O^p, \lambda) \quad (3)$$

where S is the sequence of states, S^* is the optimized sequence estimated from the Viterbi algorithms.

Next, the action generator determines actions from these estimated states as shown in Eq. (4).

$$O^{A} = \arg\max_{s} P\left(O^{a} \middle| S^{*}, \lambda\right) \quad (4)$$

where O^{A} is the stream of actions generated by the generation model.

The DHMM model with fully-connected states is implemented with 5 hidden units for each single DHMM using PMTK3 toolkit of Matlab [22]. Mihoub [5] showed that the results were not improved by using 6 or 7 unit states.

4.2. Dynamic Bayesian networks

A Dynamic Bayesian network (DBN) is a Bayesian network (BN) with variables linked by temporal dependencies. The network is a probabilistic graphical model that features the probabilistic relationships between random variables via a directed graph (DAG) in which nodes represent random variables and edges present conditional dependencies. A DBN has the ability to deal with uncertainty and to model complex temporal relationship among variables thanks to the intra-slice and inter-slice dependency structures which can be learnt from data by measuring mutual information between children and parent nodes as illustrated in Figure 3. In addition, parameters of DBN model can be also learnt by Expectation and Maximization method

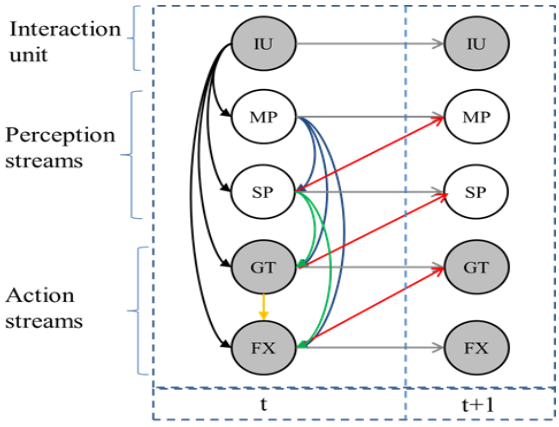


Figure 3. The learned structure of the DBN model: gray circles cue the predicted variables in the inference stage (reproduced from [6]).

The learned DBN model can be used for inference with junction tree algorithm. There are several inference methods to estimate the sequence of actions either on-line or off-line. The *filtering* inference method estimates unobserved nodes $X_t = (IU_t, GT_t, FX_t)$ of the model at time t given the sequence of observed nodes $Y_{1...t} = (SP_{1...t}, MP_{1...t})$ as shown in Eq. (5) below:

$$X_t^{*} = \arg\max_{X_t} P\left(X_t \mid Y_{1...t}\right) \quad (5)$$

The *smooth* inference method estimates the action X_t^{*} given the whole perception sequence, as given in Eq.(6)

$$X_t^{*} = \arg\max_{X_t} P\left(X_t \mid Y_{1...T}\right) \quad (6)$$

The DBN model was implemented using Bayes Net Toolbox [23] for inference and training in which the intra-slice structure and inter-slice structure were leant by K2[24] and REVEAL[25] algorithm, respectively.

4.3. Recurrent neural networks

Recently, recurrent neural network (RNN) has been applied to sequential data due to its ability to "remember" information which has been getting through. However, standard RNNs have difficulty in capturing long-term dependencies because of the vanishing problem of fixed feedback i.e. the convergence of geometric series. Long-short term memory (LSTM) RNN is able to prevent vanishing problem by including binary gates to each neuron that control whether each memory cell should process the available input, use feedback or deliver output.

Another RNN architecture is bidirectional recurrent neural network (BiRNN), which consists in combining the processing's of the same data sequence in both forward and backward direction performed by two distinct RNN. Their two output layers are then connected to one additional layer that combines the outputs once the whole sequence has been processed. BiRNN has improved the performance in many sequence learning tasks where the result can be postponed at the end of the sequence[26] [27].

4.4. Single task and multi-task

We have built discriminative multimodal interactive models using LSTM so as to improve the sensitivity of internal/latent variables to long-range structural dependencies. LSTM can be trained to directly map perception to action without considering an a priori knowledge of the underlying structure of the interaction, i.e. the interaction units (IU) introduced by Mihoub et al [Mihoub et al. 2015a] [6].

Multi-task learning [28] is meant to (1) implicitly structure the main mapping task by feeding the network with additional objectives and (2) prevent over-fitting with additional but related tasks. We also applied the multi-task methodology to implicitly structure the prediction of actions (main task) by also predicting interaction units (cognitive states/subtasks). Long-term and short-term processing capabilities of the multi-tasking LSTM are expected to benefit both to high-frequency (i.e. mapping actions) and low-frequency (i.e. recognizing units) tasks.

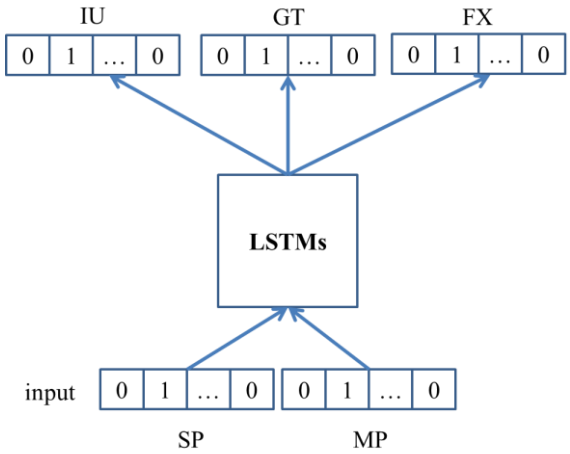
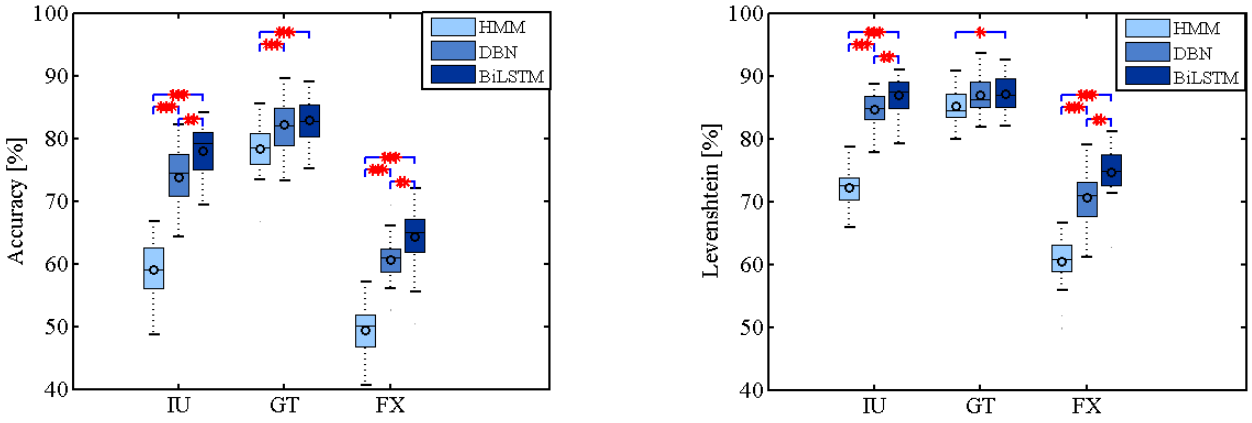


Figure 4. Schematic model of multi-tasking LSTMs. As for DBN, input streams include MP and SP. Identically, output streams are GT and FX. IU is treated as a secondary task for regularization purpose.

Figure 4 illustrates the multimodal interactive behavioral model using multi-tasking RNNs. The main task remains to predict action events (FX and GT) from the input-perception events (MP and SP) shown. The secondary task thus consists in predicting IU. The loss function of LSTM model will thus be the sum of the loss function of IU, GT and FX. Since all variables are discrete with almost identical cardinal, no weighting was performed. Neither did we decrease IU contribution as a function of iterations, because it could have been favoring the main task after convergence.



(a) Raw F-score **(b) F-score with Levenshtein alignment**

Figure 5. Offline generation: comparing performance of the joint estimation of the 3 different streams (IU, GT, FX) with the methods HMM, DBN vs. BiLSTM. (a) raw F-score, (b) F-score with relaxed alignment. The number of stars above the links between scores cue significant F-probability of Tukey post-hoc tests ('*' with p<1e-3, '**' with p<1e-2,'*' with p< 0.05). Boxes' internal lines give mean values while circles give median values of the evaluations.**

In this research, we build each multi-tasking RNN models including a hidden layer. The LSTM model has a forward LSTM with 35 gated units in the hidden layer. Otherwise, the BiLSTM model includes one forward LSTM and one backward LSTM with the same number of gated units in the hidden layer. The outputs of the two LSTMs are then connected to a time distributed dense layer applied at each time step (i.e. the output layer of the BiLSTM) with soft-max activation functions. The cardinal of the outputs of the forward and backward LSTM as well as the BiLSTM equals the sum of the cardinals of the different classifying tasks. Both of LSTM and BiLSTM model were implemented by using Keras[29].

5. Results

The LSTM and Bi-LSTM can automatically learn contextual variables from the interaction scenario. In order to compare the efficiency of the methods, actions (FX and GT) generated offline by BiLSTM are first compared with HMM [5] and DBN [6]. In addition, online predictions of the actions by LSTM are also compared with short-term Viterbi decoding of HMM and online filter prediction of DBN.

For all models, leave-one-out cross validation is applied to the 30 folded games. Both frame-by-frame comparison and Levenshtein distance estimation [30] are performed. We also perform coordination histogram, as proposed in [6], to capture global coordination patterns between different modalities given synchronous streams of discrete events. A coordination histogram computed for one modality cumulates the delays between each event in this modality and the nearest events observed in the other modalities.

5.1. *Offline task*

Figure 5b illustrates the accuracy and Levenshtein comparison for all of the methods in offline prediction tasks. Because of the direct dependency between input and output observations, the rates of DBN outperform HMM (no direct relationship between input and output) for all cues: IU (74% vs. 59%), GT (82% vs. 78%) and FX (61% vs. 49%). The BiLSTM model surpasses both other methods for IU (79%) and FX (64%) prediction (95% confidence level), respectively, while the accuracy of GT prediction caps at 83%. All prediction accuracy rates are much higher than the empirical chance levels of the tasks, i.e. 21% for IU, 34% for GT and 20% for FX. The same observations apply for the Levenshtein distance. These good results may be explained by the ability of LSTM to learn complex syntactic organization hidden in the data from the surface structure, notably causal relations that are spanning across IUs.

Figure 6 displays chronograms of input and output sequences predicted by the different models. The two first rows show the input sequences from the instructor: speech *SP* with 5 values (cube, location, reference, none, else) and arm gesture of manipulator *MP* with 4 values (rest, grasp, manipulate, end). The three final rows superimpose predictions of output streams *GT* and *FX* and *IU* in the different methods to the ground truth. Most onsets of predicted events by BiLSTM for the output streams are close to onsets observed in the ground truth, while onsets predicted by HMM are generally the most distant ones. This is confirmed by evaluating coordination histograms (see next).

Table 1. Chi squared distances between the coordination histograms of ground truth vs. those of the different off-line models. Note that degrees of freedom (df<10) depend on the distribution of delays in the different percentiles. Since events are sampled at 25Hz, the minimum bin is 40ms.

Stream	HMM	DBN	Bi-LSTM	df
SP	1054	78	**72**	**8**
GT	783	375	**122**	6
FX	1327	199	**92**	**8**

5.2. *Coordination histograms*

Coordination histograms give a global picture of the micro-coordination patterns between each modality and the other ones. These histograms proposed by Mihoub et al [6] basically collect the delays between events in one modality and the closest one in the others. Figure 7 shows coordination histogram of the methods for ground truth (first row), BiLSTM (second row), DBN (third row) and HMM (final row) corresponding to SP (first column), GT (second column) and FX (last column). Pearson's chi-squared (χ^2) distances between the histograms of the ground truth and the different models are calculated and shown in Table 1. Note that cue-specific bins are computed as 10-quantiles of the distribution of events collected by all systems. All histograms significantly differ from each other ($p<1e^{-3}$) except DBN and Bi-LSTM for SP. The smallest χ^2 distances are those of BiLSTM, which demonstrates that the BiLSTM generates the most faithful behavioral coordination patterns.

(a)

(b)

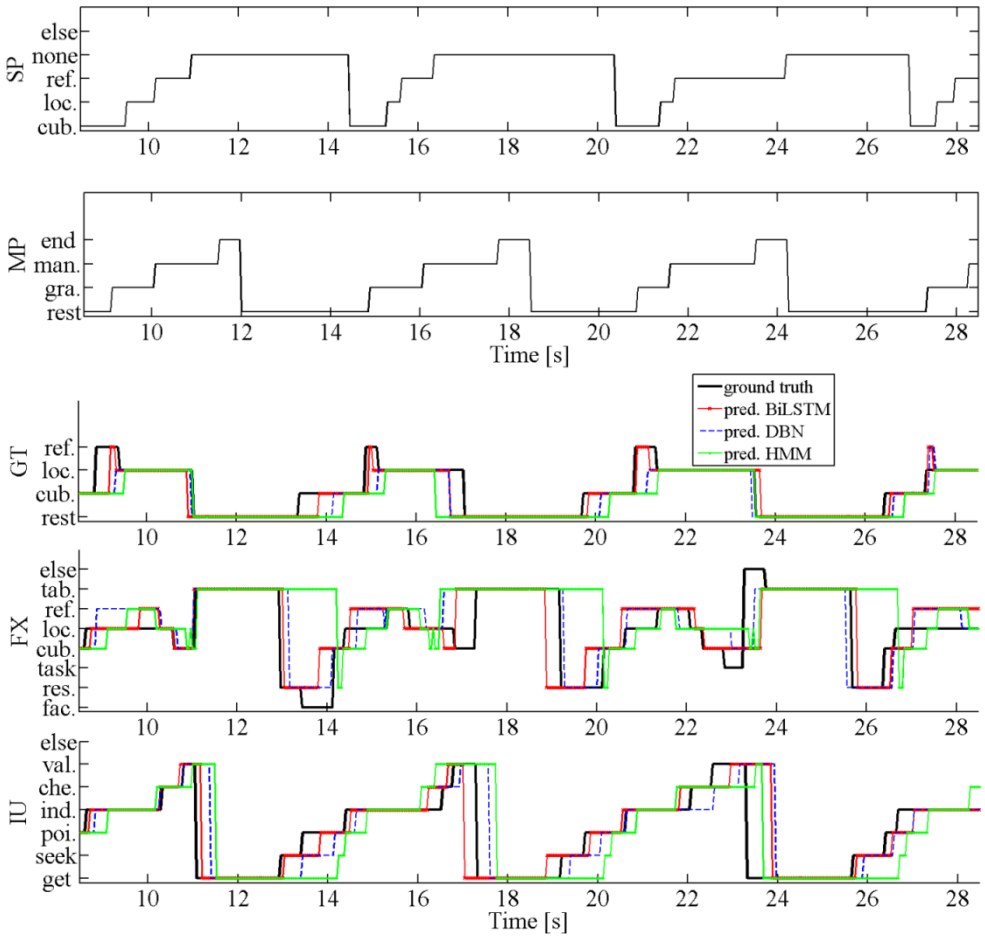


Figure 6. Input and output sequences: (a) the two top inputs MP and SP. (b) superposition of ground truth and output streams (GT, FX) and IU estimated by the different methods proposed in the paper.

5.3. Online tasks

One of the main challenges of the multimodal interactive behavioral model is to on-line drive the gesture controllers of one humanoid robot in face-to-face interaction with a human partner [31]. For this purpose, the model's output should be computed incrementally as the input sequence unveils.

Table 2. Chi-squared distances between the coordination histograms of ground truth vs. those of the different online models.

Streams	HMM	DBN	LSTM	df
SP	1114	1167	**253**	6
GT	1225	1004	**252**	4
FX	749	402	**56**	7

This section presents prediction results of the different online methods: the LSTM, the filter prediction of DBN and the HMM with Bounded short-time Viterbi.

The exact-rate and Levenshtein comparison for all of the methods in the online prediction tasks are performed and respectively shown in Figure 8a and Figure 8b. Similarly to the off-line task results, with Levenshtein estimation, DBN significantly (with 95% confident level) outperforms HMM for both IU (69.64% vs 67.64%) and FX (64.31% vs 60.97%) predictions. While the GT prediction of LSTM is almost the same as the others (84.72% for LSTM, 84.87% for DBN, 83.85% for HMM), LSTM surpasses the other methods for the prediction of IU and FX at respectively 82.93% and 70.72%.

Similarly to Table 1, Table 2 illustrates chi-squared distances between coordination histograms of the ground-truth and predictions of the three methods with the different cues. All histograms significantly differ from each other ($p<1e^{-3}$) except HMM and DBN for SP. Again, the smallest distances are those of LSTM method. These results show the effectiveness of LSTM in online prediction of faithful multimodal streams which are properly coordinated with each other.

6. Comments and discussion

The LSTM behavioral model benefits from extracting contextual information from data, instead of being limited to the boundaries of the hidden states of HMM or the immediate previous frames of the DBN dependency graph. We explored several ways to introduce latent variables in the DBN structure, notably by bootstrapping these latent variables by aligning HMM states. This does not improve DBN performance in any way. In contrast, LSTM behavioral model has the possibility to draw contextual information far away in the past history. Contextual information may in fact span large lags. For example, Richardson et al [32] have notably shown that a listener will most likely be looking at an object 2 seconds after his/her interlocutor has been paying attention to it. Mihoub et al [5] have effectively shown that adding one frame at around 2 seconds before the current input as contextual information optimally boosted HMM performance for gaze prediction from speech activity. Coordination histograms show that ground truth intermodal coordination does not exhibit fixed delays between events but a rather complex cue-dependent distribution. LSTM has the capacity to modulate memory span according to the current input

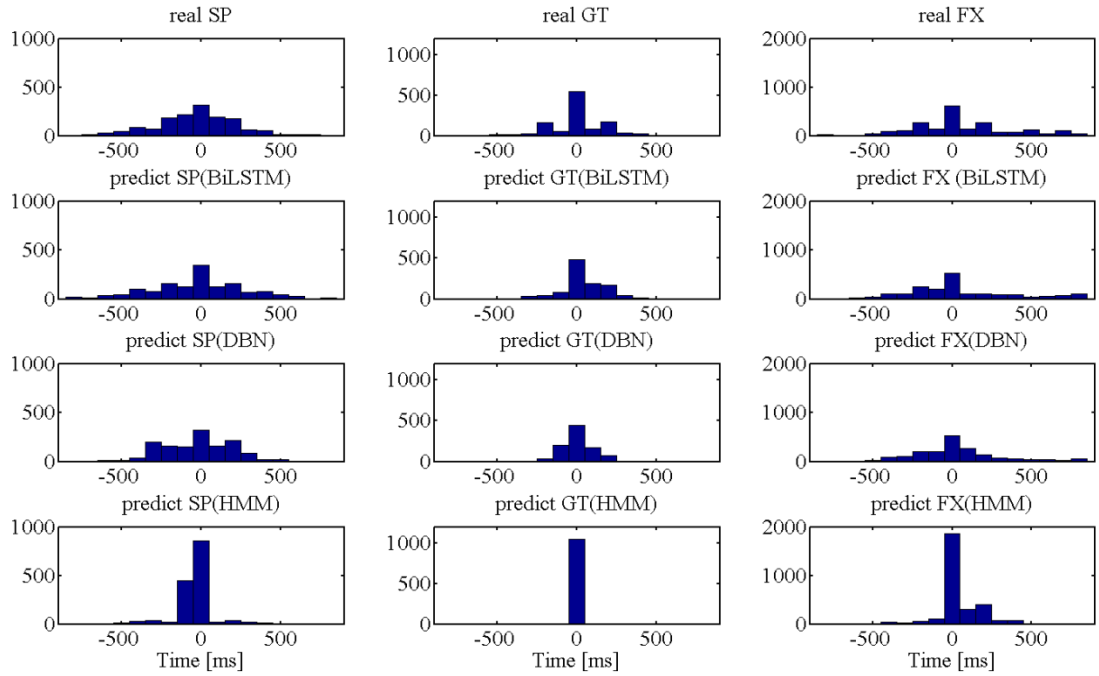
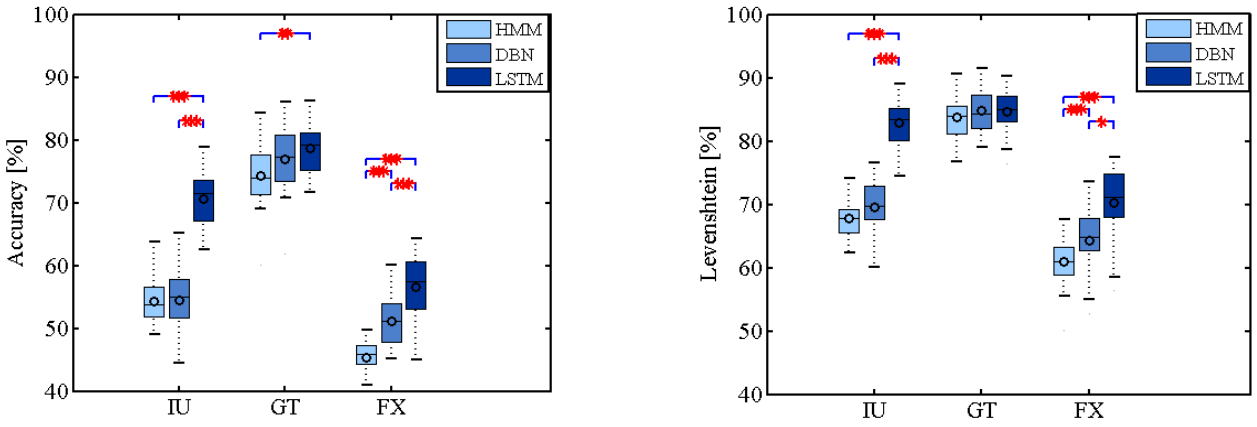


Figure 7. Comparing ground truth coordination histograms (top) with those computed with streams predicted by different offline methods, from top to bottom BiLSTM, DBN and HMM respectively. (a) speech coordination with gesture and gaze (b) gesture coordination with speech and gaze (c) gaze coordination with speech and gesture.



(a) Raw F-score

(b) Fscore with Levenshtein distance

Figure 8. Performance of the different methods for the on-line prediction tasks. Same conventions as for Figure 5

and the progress of the interaction without unnecessarily increasing the input window.

Note also that our task involves a sequence of elementary interactive skills (our IUs) with low complexity. We expect the ability of LSTM to implicitly stack features to ease the carry-over of information when the task complexity increases.

7. Conclusions & perspectives

In this paper, we present multimodal interactive behavioral models based on recurrent neural networks, namely Long-Short Term Memory (LSTM) RNN for online prediction and Bidirectional LSTM (BiLSTM) for off-line prediction. The proposed methods achieve a better performance than statistical methods with regards to both prediction performance and intermodal coordination.

In our future work, we plan to confront these models to more complex face-to-face interactions, notably neuropsychological interviews. The outputs of the behavioral models would then drive the gestural controllers of our iCub robot [31]. Subjective assessments would then be conducted to evaluate the relevance of the on-line models for the control of interactive behaviors.

The quest for performance should be moderated by the fact that models may predict alternative multimodal behaviors that are not actually observed during training but that may be appropriate acceptable variants of this particular social context. For now, human raters are the ultimate referees of the quality of interactive behaviors. We expect to extend the on-line evaluation procedure we proposed in [31] to autonomous human-robot interactions.

Acknowledgments

This research is supported by the ANR SOMBRERO (ANR-14-CE27-0014) and EQUIPEX Robotex (ANR-10-EQPX-44-01). We thank Silvain Gerber for performing the statistical analysis of coordination histograms.

References

[1] J. Cassell, H. Vilhjalmsson, M. Steedman, BEAT: the behavior expression animation toolkit, International Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, 2001: pp. 477–486.

[2] J. Lee, S.C. Marsella, Nonverbal behavior generator for embodied conversational agents, International Conference on Intelligent Virtual Agents (IVA), Marina Del Rey, CA, 2006: pp. 243–255.

[3] K. Thórisson, Natural turn-taking needs no manual: Computational theory and model, from perception to action.Multimodality in language and speech systems. Springer Netherlands, 2002. 173-207.

[4] K. Otsuka, H. Sawada, J. Yamato, Automatic Inference of Cross-modal Nonverbal Interactions in Multiparty Conversations from Gaze, Head Gestures, and Utterances "Who Responds to Whom, When, and How?," in: International Conference on Multimodal Interfaces (ICMI), Nagoya, Japan, 2007: pp. 255–262.

[5] A. Mihoub, G. Bailly, C. Wolf, Learning multimodal behavioral models for face-to-face social interaction, Journal on Multimodal User Interfaces (JMUI). 9 (2015) 195–210.

[6] A. Mihoub, G. Bailly, C. Wolf, F. Elisei, Graphical models for social behavior modeling in face-to face interaction, Pattern Recognition Letters. 74 (2016) 82–89.

[7] M. Vrigkas, C. Nikou, I.A. Kakadiaris, A review of human activity recognition methods, Frontiers in Robotics and AI. 2 (2015) paper n°28.

[8] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, M. Kankanhalli, Benchmarking a Multimodal and Multiview and Interactive Dataset for Human Action Recognition, IEEE Transactions on Cybernetics. 99 (2016).

[9] K. Noda, H. Arie, Y. Suga, T. Ogata, Multimodal integration learning of robot behavior using deep neural networks, Robotics and Autonomous Systems. 62 (2014) 721–736.

[10] D. Vogt, H.B. Amor, E. Berger, B. Jung, Learning two-person interaction models for responsive synthetic humanoids, Journal of Virtual Reality and Broadcasting. 11, paper no°1, (2014).

[11] W. De Mulder, S. Bethard, M.-F. Moens, A survey on the application of recurrent neural networks to statistical language modeling, Computer Speech & Language. 30 (2015) 61–98.

[12] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, (2014): pp. 3104–3112.

[13] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128-3137

[14] F.A. Gers, N.N. Schraudolph, J. Schmidhuber, Learning precise timing with LSTM recurrent networks, Journal of Machine Learning Research. 3 (2002) 115–143.

[15] F.J. Ordóñez, D. Roggen, Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition, Sensors. 16 (2016) 115.

[16] E. Tsironi, P. Barros, S. Wermter, Gesture Recognition with a Convolutional Long Short-Term Memory Recurrent Neural Network. Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2016: pp. *213-218*.

[17] L. Tian, J.D. Moore, C. Lai, Emotion recognition in spontaneous and acted dialogues, in: International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2015: pp. 698–704.

[18] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, Social LSTM: Human trajectory prediction in crowded spaces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: pp. 961–971.

[19] H.C. Ravichandar, A. Kumar, A. Dani, K.R. Pattipati, Learning and Predicting Sequential Tasks Using Recurrent Neural Networks and Multiple Model Filtering, AAAI Fall Symposium Series, 2016. http://www.aaai.org/ocs/index.php/FSS/FSS16/paper/viewPaper/14105 (accessed June 7, 2017).

[20] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, H. Sloetjes, Elan: a professional framework for multimodality research, Proceedings of the International Conference on Language Resources and Evaluation (LREC), 2006: pp. 1556-1559)

[21] P.P.G. Boersma, Praat, a system for doing phonetics by computer, Glot International. 5, 9/10 (2002), 341-345.

[22] M. Dunham, K. Murphy, PMTK3: Probabilistic modeling toolkit for Matlab/Octave, version 3, 2012. (2012).

[23] K.P. Murphy, others, Bayes net toolbox, Http://Www.cs.ubc.ca/Murphyk/Software/BNT/Usage.html. (2002).

[24] G.F. Cooper, E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, Machine Learning. 9 (1992) 309–347.

[25] S. Liang, S. Fuhrman, R. Somogyi, Reveal, a general reverse engineering algorithm for inference of genetic network architectures, Pacific Symposium on Biocomputing. 3 (1998) 18–29.

[26] A. Graves, N. Jaitly, A. Mohamed, Hybrid speech recognition with deep bidirectional LSTM, in: Workshop Automatic Speech Recognition and Understanding (ASRU), IEEE, 2013: pp. 273–278.

[27] R. Brueckner, B. Schulter, Social signal classification using deep BLSTM recurrent neural networks, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014: pp. 4823–4827.

[28] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Transactions on Knowledge and Data Engineering. 26 (2014) 1819–1837.

[29] F. Chollet, Keras (2015), URL Http://Keras. Io. (n.d.).

[30] L. Yujian, L. Bo, A normalized Levenshtein distance metric, IEEE Transactions on Pattern Analysis and Machine Intelligence. 29 (2007) 1091–1095.

[31] Nguyen, Duc-Canh, Bailly, Gérard, Elisei, Frédéric, Conducting neuropsychological tests with a humanoid robot: design and evaluation, in: IEEE International Conference on Cognitive Infocommunications – CogInfoCom, Wroclaw, Poland, 2016: pp. 337–342.

[32] D.C. Richardson, R. Dale, K. Shockley, Synchrony and swing in conversation: coordination, temporal dynamics, and communication, in: I. Wachsmuth, M. Lenzen, G. Knoblich (Eds.), Embodied Communication, Oxford University Press, Oxford, UK, 2008: pp. 75–93.