Learning multimodal behavioral models for face-to-face social interaction

Abstract

The aim of this paper is to model multimodal perception-action loops of human behavior in face-to-face interactions. To this end, we propose trainable behavioral models that predict the optimal actions for one specific person given others' perceived actions and the joint goals of the interlocutors. We first compare sequential models - in particular Discrete Hidden Markov Models (DHMMs) - with standard classifiers (SVMs and Decision Trees). We propose a modification of the initialization of the DHMMs in order to better capture the recurrent structure of the sensory-motor states. We show that the explicit state duration modeling by Discrete Hidden Semi Markov Models (DHSMMs) improves prediction performance. We applied these models to parallel speech and gaze data collected from interacting dyads. The challenge was to predict the gaze of one subject given the gaze of the interlocutor and the voice activity of both. For both DHMMs and DHSMMs the Short-Time Viterbi concept is used for incremental decoding and prediction. For the proposed models we evaluated objectively several properties in order to go beyond pure classification performance. Results show that Incremental DHMMs (IDHMMs) were more efficient than classic classifiers and superseded by Incremental DHSMMs (IDHSMMs). This later result emphasizes the relevance of state duration modeling.

Keywords

Sensory-motor behavior, interaction unit recognition, gaze prediction, Hidden Semi-Markov Model

Introduction

Face-to-face interaction is one of the basic elements of the human social system [1]. Nevertheless, it remains a complex and sophisticated bidirectional multimodal phenomenon in which partners continually convey, perceive, interpret and react to the other person's verbal and co-verbal displays and signals [2]. Studies on human behavior have confirmed for instance that co-verbal cues – such as body posture, arm/hand gestures, head movements, facial expressions, and eye gaze – strongly participate in the encoding and decoding of linguistic, paralinguistic and non-linguistic information. Several researchers have notably claimed that these cues are largely involved in maintaining mutual attention and social glue [3].

Human interactions are paced by multi-level perception-action loops [4]. A multimodal behavioral model aims at simulating perception for action, i.e. analyzing the scene, estimating the intended mutual goals and finally predicting and generating multimodal behavior of one target party.

Our challenge is here to train statistical multimodal behavioral models that learn by observation of the behavior of one target human subject during human-human

interactions i.e. map perception to action given exemplars of joint multimodal perceptuo-motor scores (perception-action streams of the interacting subjects). In our work, these models are trained from parallel speech and gaze data collected from interacting dyads. The challenge was to predict the gaze of one subject given the gaze of the interlocutor and the voice activity of both. The long-term goal of this research is to endow - via these behavioral models - artificial agents with social skills enabling them to engage believable interactions with human interlocutors. In this context, we present and compare two candidate models: (1) Hidden Markov Models (HMM) and their different versions which take into account an underlying structure of the causal relations between perception and action cues over time. (2) Classifiers – Support Vector Machines (SVM) and Decision Trees – that perform a direct mapping between perception and action cues without any explicit sequential modeling.

The paper is organized as follows: The first section reviews the state-of-the art of statistical analysis and generation of multimodal behaviors in face-to-face interaction. The challenging statistical models are described in details in section 2. Section 3 illustrates the application of our models on data collected in a previous experiment [5]. In section 4, all results are given and a comparison between IDHMM and SVM is made. We further introduce a specific initialization of the IDHMM model that favors the capture of temporal cycles in perception-action loops and we analyze the impact of explicit state duration modeling.

1 Related Work

This research is a part of Social Signal Processing (SSP) [6] which is a new emerging domain, spanning research not only in signal and image processing but also in social and human science involving sociology, psychology and anthropology [7]. In recent years, it has become an attractive research area and there is an increasing awareness about its technological and scientific challenges. SSP essentially deals with the analysis and the synthesis of multimodal behavior in social interactions. These are the two main tasks of a behavioral model.

Actually, automatic conversation and scene analysis [8] tries to infer information about social activities (e.g. addressing, turn taking, backchannel), social emotions (e.g. happiness, anger, fear), social relations (e.g. roles) as well as social attitudes (e.g. degree of engagement or interest, dominance) from raw social signals [6]. Several computational models have been proposed to cope with these problems. Pentland et al. [9] [10] [11] have characterized face-to-face conversations using wearable sensors. They have built a computational model based on Coupled Hidden Markov Models (CHMMs) to describe interactions between two people and predict their dynamics. The model is called the influence model. Otsuka et al. [12] proposed a Dynamic Bayesian Network (DBN) to estimate addressing and turn taking ("who responds to whom and when?"). The DBN framework is composed of three layers. The first one perceives speech and head gestures, the second layer estimates gaze patterns while the third one estimates conversations regimes. While the first layer is observable, the others are latent and should be estimated. Similarly, in order to recognize individual and group actions, Zhang

et al. [13] used a two-layered HMM. The first layer estimates individual actions from raw audiovisual data. The second one infers group actions taking into account the estimations of the first layer. For social affect detection, Petridis and Pantic [14] presented an audiovisual approach to distinguish laughter from speech (for a speaker) and showed that this approach outperforms the unimodal ones. The model uses a combination of AdaBoost and Neural Networks, where AdaBoost is used as a feature selector rather than a classifier. The model achieved a 86.9% recall rate with 76.7% precision. In the same context, ANNA (artificial neural network with a feedback loop) [15], RNN (Recurrent Neural Network) [16][17] were also used to detect social emotions integrating information from both visual and audio data streams. A Decision Tree is used in [18] for automatic role detection in multiparty conversations. Based mostly on acoustic features, the classifier assigns roles to each participant including effective participator, presenter, current information provider, and information consumer. In [19], Support Vectors Machines have been used to rate each person's dominance in multiparty interactions. The results showed that, while audio remains the most relevant modality, visual cues contribute in improving the discriminative power of the classifier. More complete reviews on models and issues related to nonverbal analysis of social interaction can be found in [20] [8] [6]. Most of the models discussed above treat only perception and scene analysis issues. The behavioral models that we propose go beyond conversation analysis, i.e. not only analyze and understand the perceived scene but also generate relevant actions from perceived cues. In a behavioral model, the analysis is done for the sake of generation: we have to keep in mind that the perception is active and that generated actions constantly modify the perception and influence interacting agents.

Actually, the generation of relevant social behaviors is the second scope of SSP. It requires robust behavioral models able of predicting the right signals to convey. One possible application is to integrate these models into social agents [21] to make them able of displaying social actions, social emotions and social attitudes through their artificial bodies. Several methods were proposed to model and synthesize human behavior. We are particularly interested on data-driven approaches which automatically infer the behavioral models from data using machine learning techniques. For instance, based on statistical gesture profiles learned from annotated multimodal behavior, Neff et al. [22] proposed to generate character-specific gesture styles capturing the individual differences of human speakers. The system takes as input arbitrary texts and produces synchronized conversational gestures in the style of a particular speaker. Admoni et Scassellati [23] introduced a preliminary model for nonverbal behavior generation. The model is aimed to be implemented in future work on a socially assistive robot for a tutoring application. Using empirical data from teachers and students in human-human tutoring interactions, the model (based on KNN algorithm) can be both predictive (recognizing the context of new nonverbal behaviors) and generative (creating new robot nonverbal behaviors based on a desired context). Only objective evaluation was performed: the model achieves 72% accuracy for gesture generation and 78% accuracy for eye gaze generation (over 4 regions of interest). For gaze generation, Lee et al. [24] implemented an eye movement model based on empirical models of saccades and statistical models of eye-tracking data. They notably observed that gaze patterns differ

depending on whether a subject is talking or listening and they included this finding in their modeling. A face character was synthesized using 3 different types of eye movements: stationary, random, and model-based. The subjective test shows that the model-generated eye movements look more natural, friendly and outgoing. Morency et al. [25] showed that sequential probabilistic models, i.e. HMMs (Hidden Markov Models) and CRFs (Conditional Random Fields), can directly estimate listener backchannels from a dataset of human-to-human interactions using multimodal output features of the speaker (spoken words, prosody and eye gaze). They notably addressed the problem of the automatic selection of relevant features and their optimal representation for probabilistic models. De kok et al. [26] presented a speaker-adaptive model to predict listener responses. The proposed approach based on a collection of individual models (trained on one interaction) outperforms the baseline model trained on all interactions. This approach first consists in identifying the closest prototypical speaker and further predicting the most adequate listener response according to the estimated communicative style. Lee et al. [27] used a probabilistic approach to predict speaker head nods and eyebrow movements for a virtual agent application. In order to learn the dynamics of head nods and the eyebrow movements, the authors explored different feature sets and different learning algorithms, namely HMM, CRF and Latent-Dynamic CRF (LDCRF). Quantitative evaluation showed that the LDCRF models achieved the best performance, underlying the importance of learning the dynamics between different gesture classes and the hidden internal orchestration of the gestures. Huang et al. [28] explored how a learning-based approach meant to model multimodal behaviors might address the limitations of heuristic-based models. They used Dynamic Bayesian Networks (DBNs) to model the coordination of speech, gaze, and gesture behaviors in narration. The evaluation of this model shows that this learning-based approach achieves similar performance compared to conventional rule-based approaches while reducing the effort involved in identifying hidden behavioral patterns. Other models using learning approaches can be found in [29][30][31][32]. More generally, these learning approaches frequently use probabilistic graphical models because of their capacity to provide a probabilistic representation of the dynamics of human behavior and to model complex multimodal relationships under uncertainty.

In the next section we will present our statistical data-driven behavioral models. Compared to scripted behaviors, they have many advantages: first, they rely on machine learning and statistical modeling to intrinsically couple perception for comprehension and perception for action, i.e. combine different scopes/windows of interaction analysis [33] and organize sequences of percepts and actions into so-called joint sensory-motor behaviors. Secondly, both analysis and generation are done incrementally which make these models usable for online applications. Thirdly, we show below that our models capture the micro-structure and some regularities of the joint behaviors that often escape to human expertise.

2 Modeling sensory-motor behaviors

We model each situated human/human - and thus human/machine - interaction into a sequence of interaction units (IUs) [34]. The number, the extent and the ordering of

these IUs depend on the task. The sequencing of the IUs, i.e. their syntax, provides a sort of behavioral grammar that chains elementary sensory-motor behaviors. A given IU can be seen as an instance of the joint cognitive states of the interacting dyads such as “thinking”, “informing”, “listening”, “taking turn”, “glazing over”, etc. A similar concept was used in [35] where the authors propose a gaze model (based on the well-known Rickel model [36][37]) driven by the cognitive operations of a virtual agent. The Rickel model assumes that gaze is closely tied to the agent’s cognitive operations and critical gaze events are used to segment the interaction into relevant IUs. These cognitive operations may include perceptions of events, the update of beliefs, understanding speech, planning, taking etc. In our modeling framework, an interaction unit (IU) further conditions the sensory-motor coupling between partners of the conversation: it contextualizes the way partners mutually signal their willingness to initiate, regulate, give, accept, acknowledge or terminate information exchanges. The achievement of an elementary interaction unit may require the sequencing of several sensory-motor states (such as mutual gaze, nodding, lip reading, beat gestures, etc.), properly coordinating the joint multimodal score and sometimes repeated over time – i.e. cycling – to secure the information exchange via multimodal complementarity and redundancy.

In this section, we present statistical/probabilistic approaches for modeling joint multimodal sensory-motor behaviors. These models should be able for a target subject (1) to estimate the interaction unit (IU) from perceptual observations (e.g. speech activity/gaze fixations of the partner); note that, when the two partners cooperate, this IU should ideally reflect the shared mental state of the conversation partners at that particular moment; (2) to generate suitable actions (e.g. his own gaze fixations, hand gestures or head movements) that reflect the current IU and his current awareness of the evolution of the shared plan.

As matter of fact, we chose HMMs [38] [39] because they have intrinsic sequential and temporal modeling capabilities. We compare here their performance with those of two well-known powerful classifiers, namely SVMs and Decision Trees. We also investigate the contribution of HSMMs compared to simple HMMs.

2.1 Incremental Discrete Hidden Markov Model (IDHMM)

As introduced above, an interaction can be seen as a sequence of discrete tasks, sub-tasks or activities [40]. Thus, in the following, we will consider a situated conversation as a sequence of interaction units (IUs) that structure the joint behavior of the conversation partners. In our model, we suppose to chain P interaction units; each IU is modeled by a single Discrete Hidden Markov Model ($\lambda_p = (A_p, B_p, \pi_p)_{p=1..P}$) whose n_p hidden states (here sensory-motor states or SMS) model the micro-syntax of IU-specific co-variations of the partners’ behaviors. The proper chaining of these HMM obeys a task-specific syntax and results from lawful mutual attention and collaborative actions. We here consider discrete observations, i.e. observations that can take on one of K possible outcomes such as gaze fixations over K regions of interest of the visual field, iconic gestures or speech over a finite vocabulary, etc.

Hence, the whole interaction is modeled by a global Discrete HMM ($\lambda = (A, B, \pi)$) that concatenates the different elementary IU-specific models (Figure 1). The global DHMM λ is composed of N hidden SMS ($N=\sum_{p=1}^{P} n_p$). As mentioned before, HMM states are associated with homogenous joint sensory-motor behaviors: given T as the length of the sequence, the observation vector $O = (o_t)_{t=1..T}$ is in fact composed of two streams: (1) the sensory stream $O^p = (o_t^p)_{t=1..T}$ collects perceptual cues; (2) the motor stream $O^a = (o_t^a)_{t=1..T}$ is responsible for initiating actions. The observation vector is then defined as follows:

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

Note that the sensory stream may include sensory consequences of self-generated actions. These may be of different natures: efferent copies of actions, proprioceptive or exteroceptive signals (such as involved in oculomotor coordination). Moreover, our SMS (Figure 1) intrinsically associate percepts and actions and may collect actual or expected responses of a self-generated action as well as motor responses for a perceived event that are appropriate to the current IU. Conversely, the discrete sensory observations can also include unknown cues: visual cues such as identity of an unknown object or agent, the gaze direction of an interlocutor are only available when the target agent is effectively gazing to them. One side benefit of joint sensory-motor learning is active perception, i.e. action for directing attention and triggering perceptual analysis of the region of interest of the audiovisual field.

2.1.1 Training recognition and generation models

Given labeled sensory-motor data, the training of IU-specific models is quite straight-forward and can be done using the classical Expectation-Maximization (EM) algorithm. The global transition matrix *A* is built from the different trained intra-HMM transitions matrices ($(A_p)_{p=1..P}$). In addition, the inter-HMMs transition probabilities are trained in order to complete this matrix *A*. Note that more sophisticated syntactic models such as n-grams can be used.

In practice, at an instant *t*, only perceptual information is available and actions have to be generated according to these input cues. For that reason, once we get the global trained HMM, two models are extracted:

- λ_R is a recognition model that selects only perception streams (i.e. $o_t = o_t^p$). λ_R performs the SMS alignment S^* with these percepts:

$$S^* = \underset{S}{\operatorname{argmax}}\ P(S|O^p, \lambda_R) \quad (2)$$

- λ_G is a generative model that samples the emission matrix *B* with action emission probabilities (i.e. $o_t = o_t^a$). Given only the SMS alignment (S^*) performed by λ_R, λ_G further generates the adequate actions.

Selection and sequencing of our SMS is solved by the process of HMM states decoding [38], usually performed by the Viterbi algorithm which normally runs in an offline mode, since it requires to entire sequence as input. We therefore perform online decoding with

an alternative approach known as the Short-Time Viterbi (STV), which also allows to control latency. It consists of using an expanding window and comparing partial paths converging to the same trajectory [41] [42] [43]. In fact, the central idea of this algorithm [42] and its variants is that the window is continuously expanding forward until a convergence/fusion point is found. When this is the case, the Viterbi algorithm is reinitialized from that point. The main advantage of this method is that the solution is strictly equivalent to the full Viterbi algorithm. The major drawback is that the fusion point can be very far ahead. In this paper, we adopted a bounded version of the STV (BSTV): we set up a threshold beyond which the path with maximum likelihood up to a given number of frames ahead of the current frame is retained when there is no fusion point within that horizon. Although the optimal solution is not always selected, the latency is fully controlled. We will show that, in our data, very short latencies do not significantly degrade the performance of the decoder.

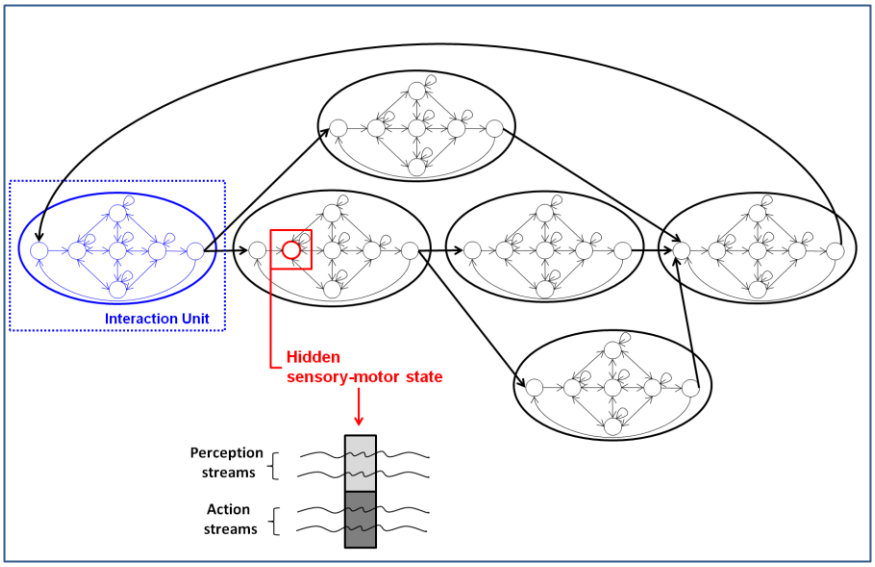


Figure 1: Management of perception-action loops in a probabilistic scheme linking observations, sensory-motor states, IUs and task syntax (sequence of IUs). Red Arrow figures here probability density functions, notably the *emission probabilities* governing the distribution of the observed frames at a particular time given the hidden state at that time. We figured here perceptual observations with light gray and motor observations with dark gray. Note that observations can combine frames sampled at the current time as well as in the past.

2.2 Incremental Discrete Hidden Semi-Markov Model

A major limitation of conventional HMMs is state duration modeling. Durations of hidden states implicitly follow a geometric distribution which may be inadequate for most applications. As an extension of the HMM, the Hidden Semi-Markov Model (HSMM) explicitly models the duration or residence time for each state [44]. The HSMM is also known as explicit duration HMM [45][38], variable-duration HMM [46], generalized HMM [47] and segmental HMM [48]. They have been successfully applied in a wide range of domains such as speech processing [49] [50], image and video analysis [51] [52], robotics [53], networks security [54], biology [47][55] and financial time series [56]. A good review of HSMMs can be found in [44]. Several approaches have been introduced to solve the problem of learning and inference in HSMM, including the approach described in the Rabiner paper [38] which was initially proposed by Ferguson [45] and then improved by Levinson [46] and Mitchell [57]. This later approach uses a Forward-Backward algorithm that estimates the joint probability that a state ends at a given time and models a series of observations until that time. A more efficient Forward-Backward algorithm with a lower complexity in time and memory was proposed in 2003 by Yu and Kobayashi [58]. In our work, we use a recent version of that algorithm [59] and the code provided by the authors.

2.3 SVMs and Decision Trees

In order to compare the IDHMM model to a baseline system, we propose to apply standard classifiers to our estimation problem, in particular SVMs and Decision Trees. Both SVMs and decision trees are among the most used and powerful classifiers. In our context, each HMM model will be compared to the outputs of two distinct classifiers trained with the same data: the first one will estimate the most likely interaction unit from perceptual observations while the second one will directly determine the most likely actions from perceptual observations. The performance of the classifiers with regards of these two tasks will be compared to the performance of the IDHMM recognizer λ_R and generator λ_G. In section 3 we will present how we applied all proposed models to our dataset.

3 Experimental setting

We used the dataset of Bailly et al. [5] who collected speech and gaze data from dyads playing a speech game via a computer-mediated communication system that enabled eye contact and dual eye tracking. The experimental setting is shown in Figure 2: the gaze fixations of each subject over 5 regions of interest (ROI: face, left & right eye, mouth, elsewhere) are estimated by positioning dispersion ellipsis on fixation points gathered for each experiment after compensating for head movements. The speech game involved an instructor who reads and utters a sentence that the other subject (respondent) should repeat immediately in a single attempt. The quality of the repetition is rated by the instructor. Dyads exchange Semantically Unpredictable Sentences (SUS) that force the listeners to be highly attentive to the audiovisual signals, notably to lip-read when listening to unknown linguistic content. The experiment was

designed to study adaptation: one female speaker called “LN” interacted with ten subjects (colleagues of roughly the same age and social status and students) both as an instructor for ten sentences and as a respondent for another set of ten sentences.

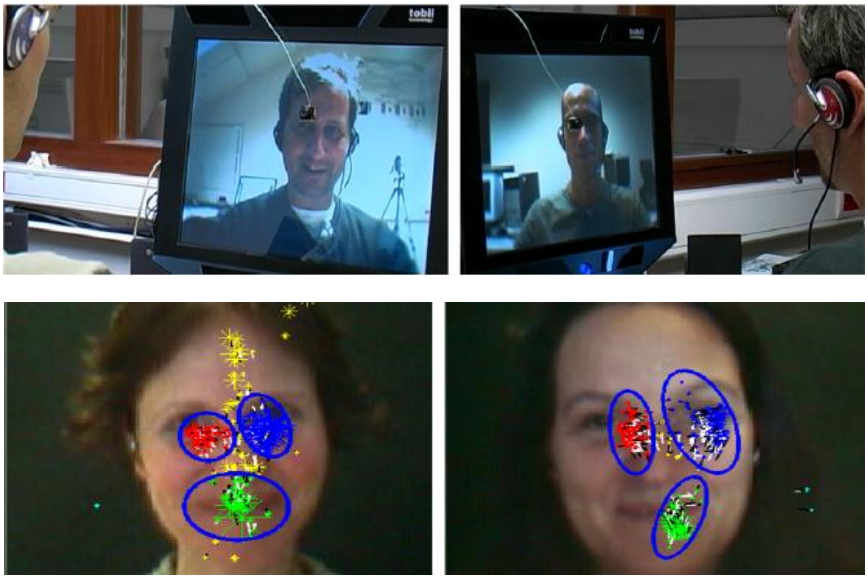


Figure 2: Mediated face-to-face conversation [5]. Top: People sit in two different rooms and dialog through couples of cameras, screens, microphones and loudspeakers. Gaze of both interlocutors are monitored by two eye-trackers embedded in the TFT screens. Note that pinhole cameras and seats are positioned at the beginning of the interaction so that the cameras coincide with the top of the nose of each partner’s face. Bottom: four regions of fixation are tracked on each speaker’s face: left and right eye, mouth and face (mainly the nose ridge).

3.1 IDHMM and classifiers

For each dyad, we have two observations streams: voice activity (v1/v2 with binary values: on/off) and gaze fixations (g1/g2 towards 5 ROI: face/mouth/left eye/right eye/else) of both speakers. Seven interaction units (IUs) [34][60] have been labeled semi-automatically (‘Read’, ‘Prephon’, ‘Speak’, ‘Wait, ‘Listen’, ‘Think’ and ‘Else’). The task syntax controlling the seven IUs is illustrated in Figure 3. We tested the ability of IDHMMs to estimate the IU for the main subject "LN" given her voice activity (v1) as well as the voice activity (v2) and gaze (g2) of her conversational partner, and then predict her own gaze behavior (g1). Consequently, we use the recognition model λ_R to decode $o_t = (v1, v2, g2)$ and next λ_G to generate her gaze (g1). The number of hidden states was set to 5 per IU resulting in 35 hidden states for the whole IDHMM. Topologies with a fixed number of 4 and 6 hidden states per IU were also tested but resulted in no significant difference in performance. The rest of parameters (i.e.(A, B, π)) were initialized randomly. For SVMs and Decisions Trees, a first classifier is used to estimate the IUs from (v1, v2, g2). Then a second classifier is used to estimate the gaze (g1) from the same data. DHMMs are trained with HTK [61], the IDHMM model was implemented in Matlab using PMTK3 toolkit [62]. For SVMs/Decision Trees, the Weka java package [63] has been used for both training and testing. For all models, 10-fold cross validation was applied.

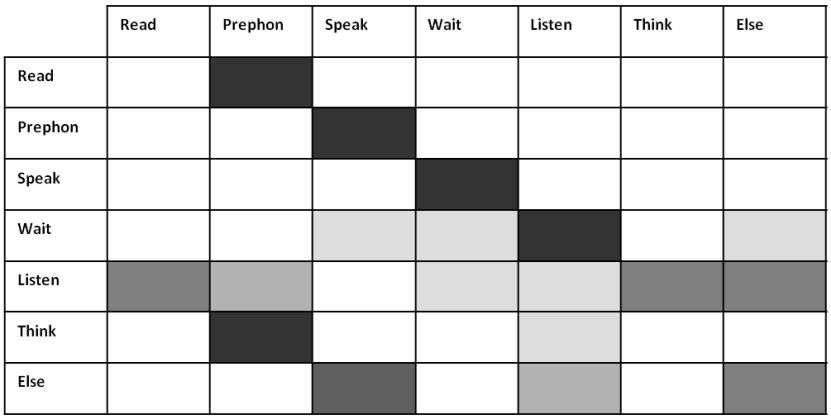


Figure 3: Task syntax: transition matrix between interaction units, darker the color is, higher is the transition probability

3.2 Data-driven dimensioning and initialization of the IDHMM model

The internal cognitive process presumably differs from one interaction unit to another. The IU "Listen" will be surely longer in time and more complex in structure than the IU "Prephon". We therefore varied the number of hidden SMS per IU according to some objective criterion. According to observations' distribution, we propose a methodology that allows us to automatically (1) select the adequate number of hidden sensory-motor states representing a given IU and (2) initialize the emission probabilities for each selected hidden state. In the following we consider three assumptions:

1. Each observation is associated with a unique sensory-motor state
2. Observations must have a significant weight in the distribution
3. The number of states to select must be in a reasonable interval

These assumptions are not strict. They allow us to initialize model parameters to semantically meaningful values. However during EM training, the learned parameters may violate these assumptions. The first assumption which uniquely relates observations to hidden SMS only initializes emission probabilities. The EM training algorithm may of course reconsider these initial distributions of emission and transition probabilities as well as IU boundaries.

We first collected all observations of a given IU and rank these observations in a descending order according to their frequency of occurrence. We then select a number of observations (SMS) whose cumulative frequency represents at least 90% of the total ground truth (assumption 2). We limit the range of SMS to 5-10 in order to preserve the relevance of the model (assumption 3). In Figure 4 we show an example of our method to determine the appropriate number of hidden states for the IU Listen. The figure shows that 15 states should be selected in order to get 90% of the cumulative frequency. With 10 hidden states selected, we retain 76% of the information. Table 1 gives the distribution of observations for the first three selected states for the IU Listen. Table 2 shows the number of SMS selected per IU and their corresponding cumulative frequency. The new IDHMM model contains 59 hidden states and emission probabilities of each state were initialized in a manner to reflect the observed distributions. The rest of training stage is similar to the first model.

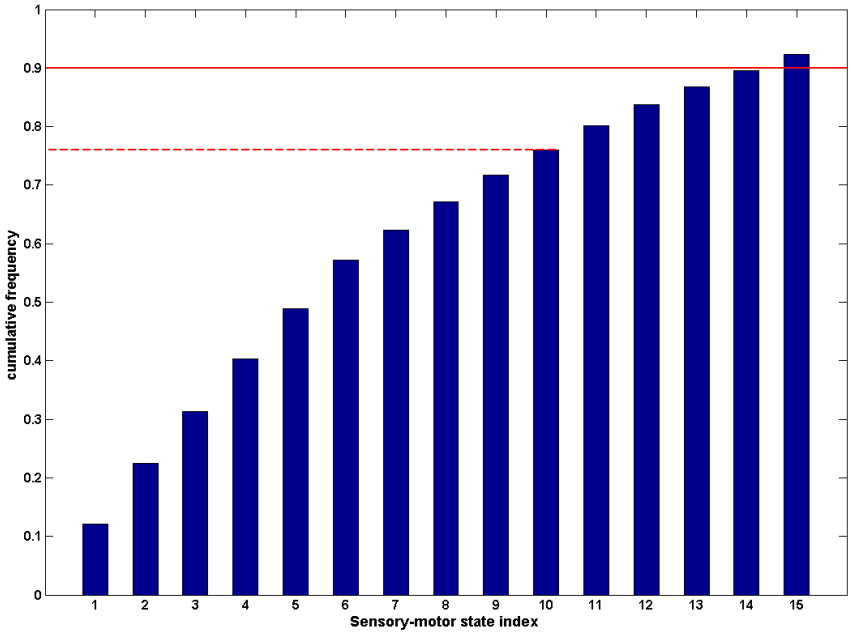


Figure 4: An example of our method to determine the number of sensory motor states. In this case the maximum number (10) SMS are selected for the IU *Listen*.

	State1	State2	State3
Gaze of the interlocutor	left eye	left eye	left eye
Speech of the interlocutor	on	on	on
Gaze of the principal subject LN	left eye	mouth	right eye
Speech of the principal subject LN	off	off	off

Table 1: The first three selected sensory-motor states for IU Listen and the dominant observations with which they have been initialized

	Number	Weight
Read	5	99%
Prephon	8	91%
Speak	10	89%
Wait	9	91%
Listen	10	76%
Think	10	85%
Else	7	90%

Table 2: The number of SMS per IU and their corresponding cumulative frequencies

3.3 The IDHSMM model

As will be shown in section 4, the data-driven dimensioning and initialization of the IDHMM model has some interesting properties, notably for capturing SMS cycles. For that reason, we chose to keep the same structure for the IDHSMM: our model contains 59 hidden states and was initialized in the same way as described in the previous subsection. Compared to HMM, the HSMM needs an additional matrix $\mathbf{D}$, in which each line describes a discrete distribution of the duration of a sensory-motor state. $\mathbf{D}$ is defined as follows:

$$D = \{d_{ij}, i = 1..N \; and \; j = 1..M\} \quad (3)$$

Where N is the number of hidden states (59 sensory-motor states), M is the maximum duration in frames and $\sum_{j=1}^{M}(d_{ij})_{i=1..N} = 1$. In our training set, M is equal to 101 frames (1 frame corresponds to 40ms). The matrix D is computed empirically from data by counting for each state *i* and duration *j* the number of occurrences of consecutive characteristic observations. For the rest of parameters, there is no difference in the training process compared to HMM. Thus, the HSMM is fully specified by $\lambda = (\mathrm{A}, \mathrm{B}, \mathrm{D}, \pi)$. Based on the concept of fusion point [42], the Forward-Backward algorithm [59] was modified in order to do incremental recognition and generation. In the next section we will discuss the results of all proposed models.

4 Results

We will start by presenting in detail the results of the baseline IDHMM model and compare its performance with classifiers. Second, the properties resulting from the data-driven dimensioning and initialization of IDHMM are discussed. Finally, by comparing HMM with HSMM performance, we will demonstrate the relevance of state duration modeling.

4.1 IDHMM results

Classification accuracy is used to evaluate interaction unit recognition, whereas the Levenshtein distance [64] is adopted for the evaluation of gaze generation because it provides adequate structural – less sensitive to fine alignments – comparison between generated and original signals. In fact, the Levenshtein distance is a metric for measuring the difference between two sequences: using dynamic time warping, it computes the minimum number of elementary operations (insertions, deletions and substitutions) required to change one sequence into the other. From this optimal alignment, recall, precision and their harmonic mean (the F-measure) can be directly computed. In this paper, all generation rates represent F-measures.

The mean recognition rate of 92% shows that STV is able to capture the structure of the interaction (see Figure 5 and Figure 9). An offline processing with an infinite horizon gives the same result which confirms the efficiency of STV performance. However, the problem with STV is mastering the output delay. We observe that 80% of latencies are less than 5 frames. But maximum values can be very high: in our case, for all subjects, the maximum latency was 259 frames which represents an unsuitable delay for real-time application. BSTV is used to limit these delays. Theoretically, an optimal trade-off ought to be sought because of the inverse relationship between performance and latency. From Figure 5 and Figure 9, we can see that our IDHMM is able to approximate the Viterbi path with low thresholds/latencies at the expense of a small degradation of IU recognition (i.e. 89% for a threshold equal to 1 frame). Moreover the mean generation performance (59%) is not affected and remains practically the same at all thresholds. The IDHMM model with one frame delay has a rate of 89% for interaction unit detection and 59% for eye gaze generation. This performance is mainly due to the strong proportions of very short latencies: deviations from the global optimal path rapidly reconnect when robust cues are encountered. Another important factor is the constrained syntax of the task: the chaining of sub-tasks is very regular and highly constraints the alignment of interaction units.

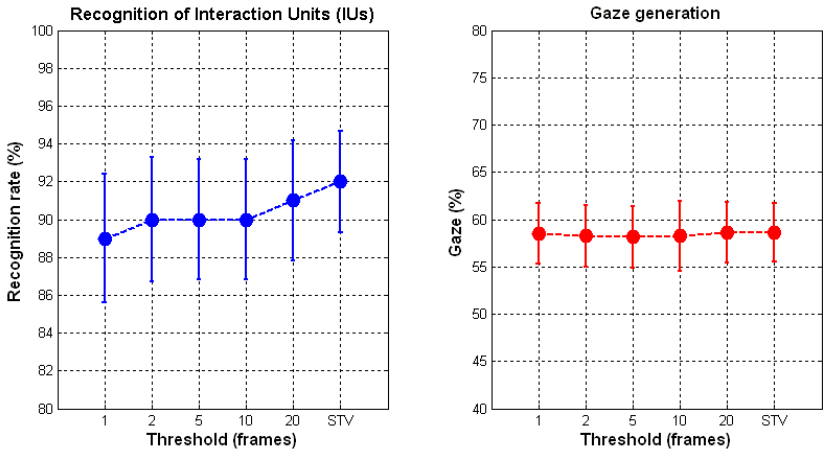


Figure 5: Recognition and generation results using IDHMM as function of threshold/latency (expressed in frames, 1 frame is 1/25th of a second)

4.2 Analyzing speaker-dependent models

Until this point, the IDHMM is interlocutor-independent (II): for each interlocutor, the corresponding II model is trained on the other 9 interactions. We build also interlocutor-dependent (ID) models [65] [26]: a set of 10 ID models is built using data from each dyad. Mirroring the training of II models, each ID model is thus trained on one interaction and tested on the 9 remaining ones. The results shows that the II models result in better performances compared with ID models: the mean behavior outperforms all individual ones. This is not surprising since II models are trained on more interactive data. Nevertheless, ID models were further used to study social proximity and relation between the subjects. Actually a multidimensional scaling (MDS) analysis based on Kruskal's normalized STRESS1 criterion was performed on ID interaction unit recognition and gaze prediction errors (see Figure 6). This analysis of the resulting proximity between ID behaviors nicely mirrors known social relationships between our target speaker LN and her interlocutors: LN consistently behaves differently when interacting with colleagues compared to students. The MDS finely reflects social distance: the PhD student supervised by LN is closer than others, junior colleagues with permanent positions are positioned far more than PostDocs. As already evidenced by Pentland and colleagues [10] [11], joint behavioral models may capture subtle adaptive cues that signal pre-existing or developing social relations. Indeed gaze is a very social signal and no doubt that social determinants of interaction such as personalities and dominance relations are mirrored in gaze behaviors: such by-product of modeling deserves further research. Note also that interactive models may capture subtle behavioral idiosyncrasies that are difficult to characterize with short-term signal-based methods that do not exploit the structure nor the intend of the interaction.

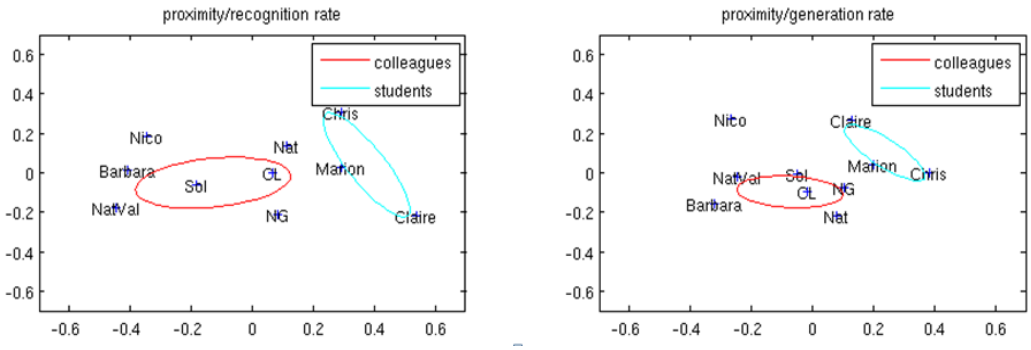


Figure 6: An MDS projection of the performances of the ID models cues proximities between interlocutor-specific behaviors: note its coherence with the a priori clustering of their social relations with "LN"

4.3 Comparison with classifiers

In this paragraph, we compare our sequential IDHMM with SVMs and decisions trees that do not explicitly perform sequential modeling. Figure 8 clearly shows that there is no significant difference between the two non-sequential classifiers. However, the IDHMM model (threshold=1) outperforms the two classifiers and the improvement provided by this model is quite significant ($p<0.05$). Moreover, Figure 9 shows that the IDHMM model is more efficient in detecting the structure of the interaction. We can see that the estimated path of interaction units correctly reflects the predefined syntax of the task. In comparison, the SVMs has more difficulty in capturing the organization of the real path (see Figure 9) and discards short interaction units: we can see that the estimated IUs are principally « Speak », « Wait » and « Listen ». This is in not in contradiction with the 81% recognition rate because these three interaction units alone represent 85% of the ground truth. This performance gap is mainly due to the sequential constraints imposed by HMMs. This lack of sequential organization impairs the performance of SVMs and Decision Trees that should exclusively exploit instantaneous information provided by the observations.

4.4 Adding contextual observations

In a previous work [66], classifiers' performances were improved by adding memory (historical values) to each observation. In fact, at a time t, the classical models only use data of that moment. In this new configuration, we added the same three attributes (*v1,v2,g2*) but from a previous instant t-T, where T is an offset parameter which was empirically chosen as T=55 frames (~ 2 seconds, see Figure 7). This optimal delay corresponds exactly to the one reported in [67] in which authors demonstrate that, if a speaker looks at an object, 2 seconds after the listener will most likely be looking at the same object. SVMs with contextual attributes lead to a generation rate of 59% (vs. 50% in the classic model). Hence, supplying the SVM model with memory relatively addressed the missing sequential aspect.

4.5 Results of IDHMM with data-driven dimensioning and initialization

The objective comparison between automatic and data-driven initialization of IDHMM does not result into significant difference in IU estimation and gaze generation. However we notice significant differences in the estimated SMS alignment paths. In Figure 10 (see

selected zones), we can see that, when transition probabilities B_p are randomly initialized, the HTK training algorithm ends roughly with a left-right HMM typology. However, when initialized with a data-driven counting procedure, the internal structure of IU becomes fully connected and exhibits SMS cycles. Figure 11 compares the number of such cycles found in ground truth data in comparison with SMS scores computed by the randomly initialized IDHMM and the data-driven initialized IDHMM. The number of cycles between sensory-motor states for the modified IDHMM is closer to ground truth's number than the baseline IDHMM model. Due to this interesting propriety, we kept the same structure for the IDHSMM model.

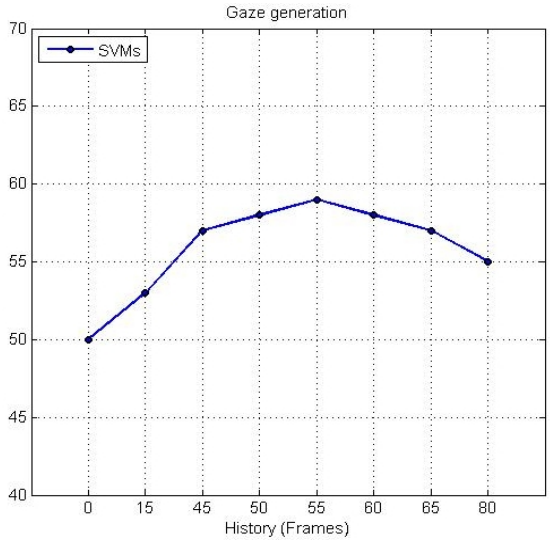


Figure 7: Optimal memory instant for SVMs. A frame equals to 40ms.

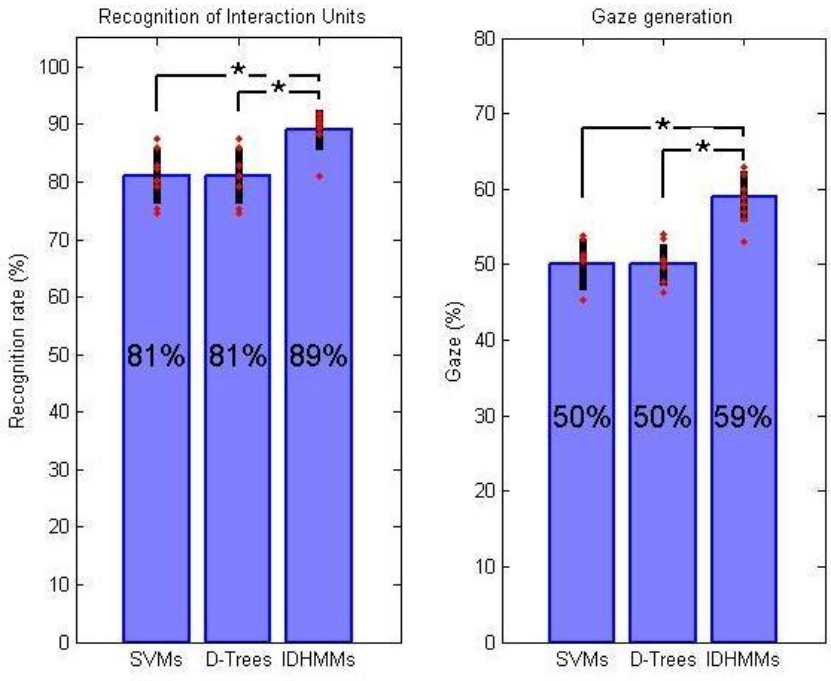
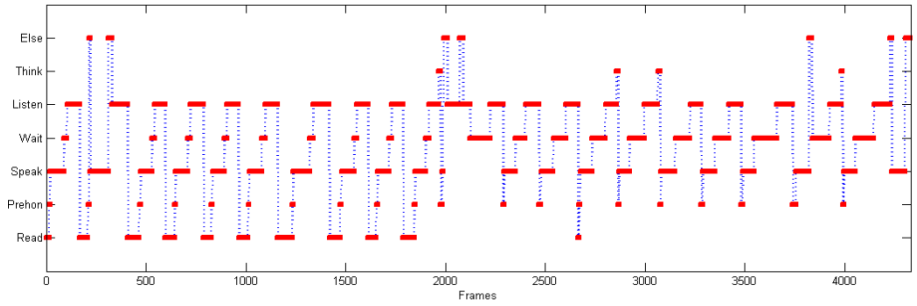
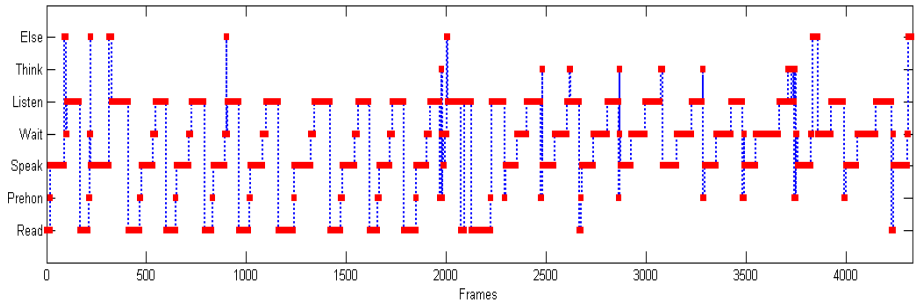


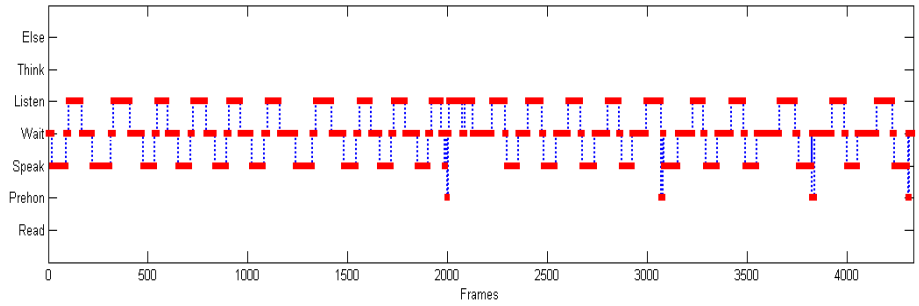
Figure 8: Results of the three models: SVMs, Decision Trees and IDHMMs



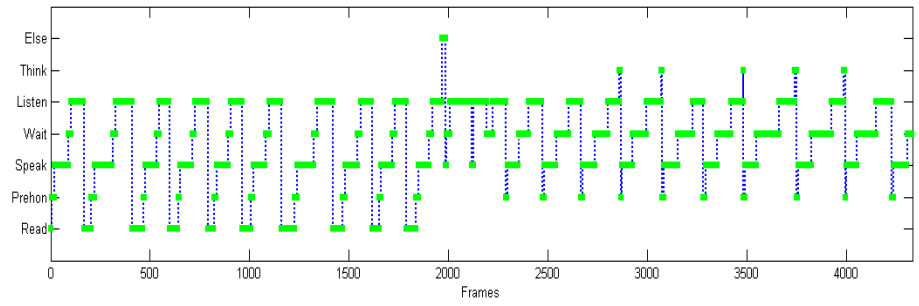
(a)



(b)

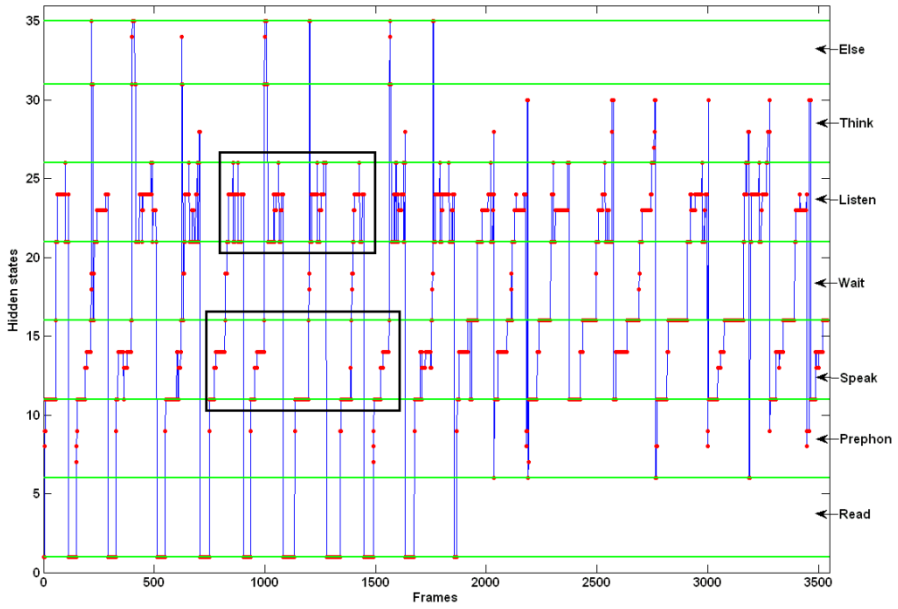


(c)

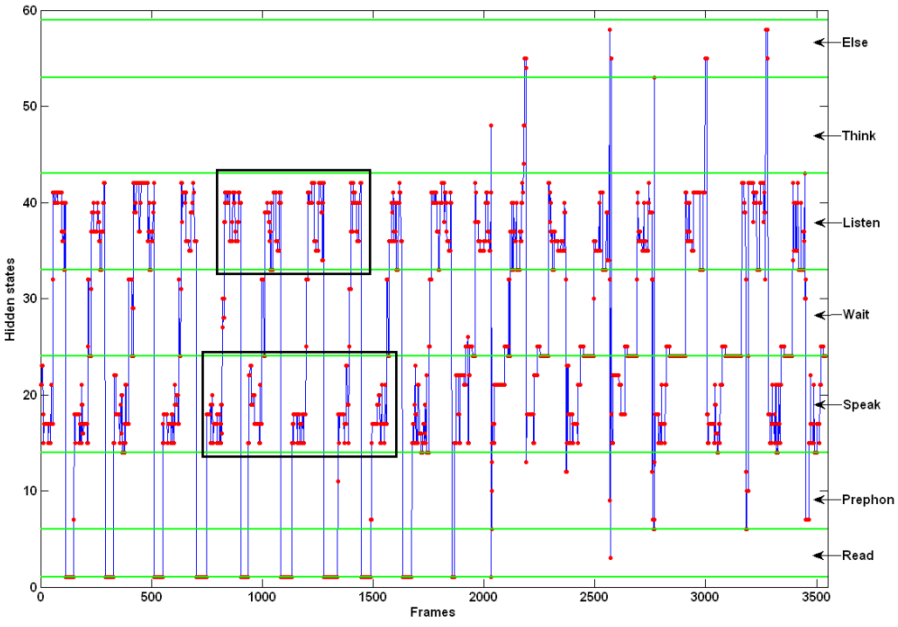


(d)

Figure 9: Estimation of the interaction units (IUs) for a specific subject (a) using IDHMM (no threshold) (b) using IDHMM (threshold=1) (c) using SVMs (d) the real IU path



(a)



(b)

Figure 10: Example of an estimated path (a) in the baseline IDHMM and (b) in the modified IDHMM

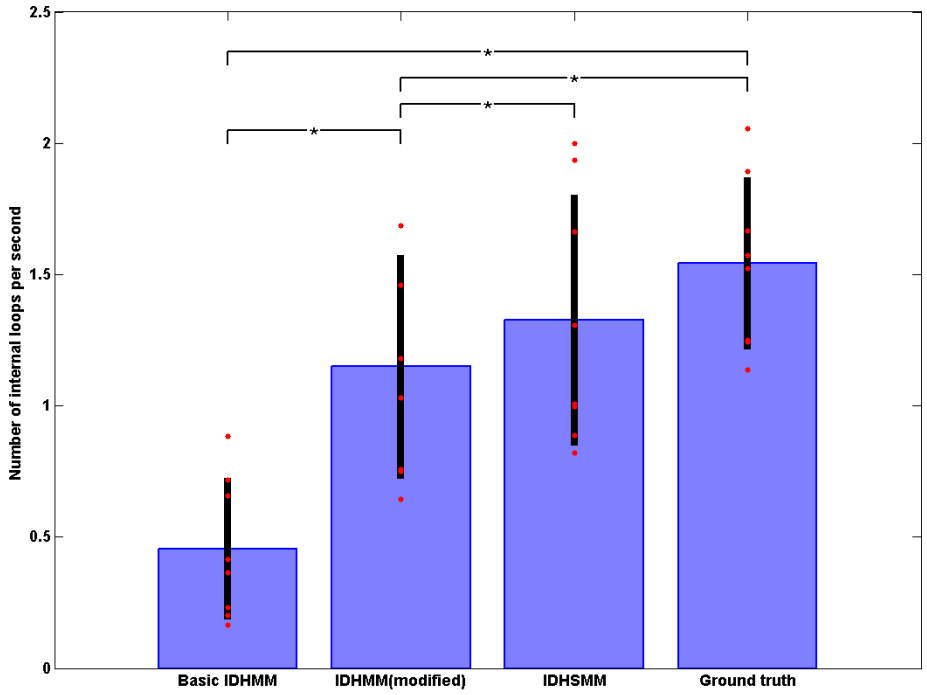


Figure 11: Number of internal cycles per second performed by sensory-motor states for all models

4.6 IDHSMM results

As argued before, state duration in classic HMMs is modeled by a geometric distribution which is not appropriate for most of physical signals [38]. We study here the capacity of our IDHSMM to overcome such limitations. General comparison results are shown in Figure 12. As we can see, recognition rates are almost the same for IDHSMM and IDHMM (90% vs. 89%). However, generation rates are significantly better for IDHSMM (63% vs. 59%). Note that this improvement stands also for non incremental processing: HSMM also outperforms HMM for generation (64% vs. 59%). The better performance in gaze generation is due to the capacity of IDHSMM to circumvent the duration of generated fixations. In fact, Figure 13 displays mean durations of each region of interest (face, right eye, etc) generated by IDHMM and IDHSMM. Compared to ground truth, the mean durations of gaze generated by IDHSMM are better than those of IDHMM in four of the five regions of interest. Another important result is that IDHSMM better captures sensory-motor cycles (see Figure 11). As a conclusion, IDHSMM generates more accurate SMS durations and then leads to more relevant motor scores.

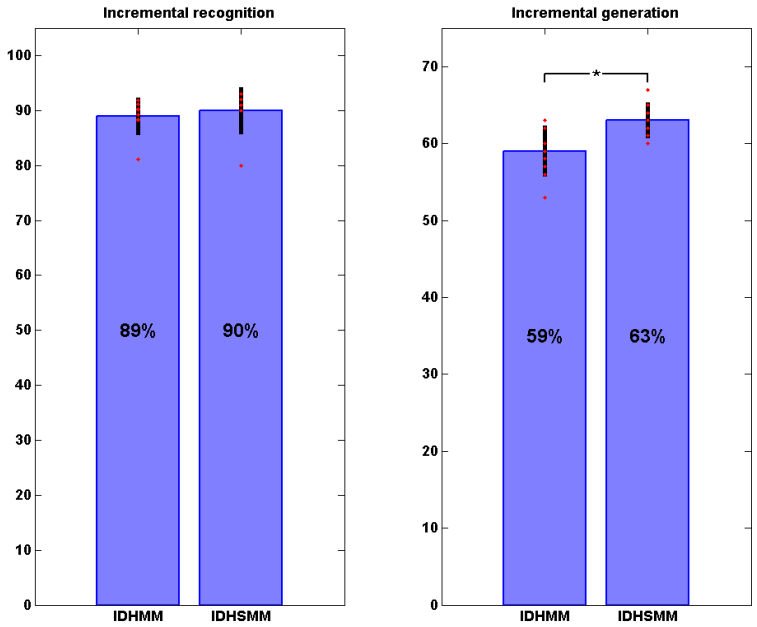


Figure 12: Comparison results between IDHMM and IDHSMM

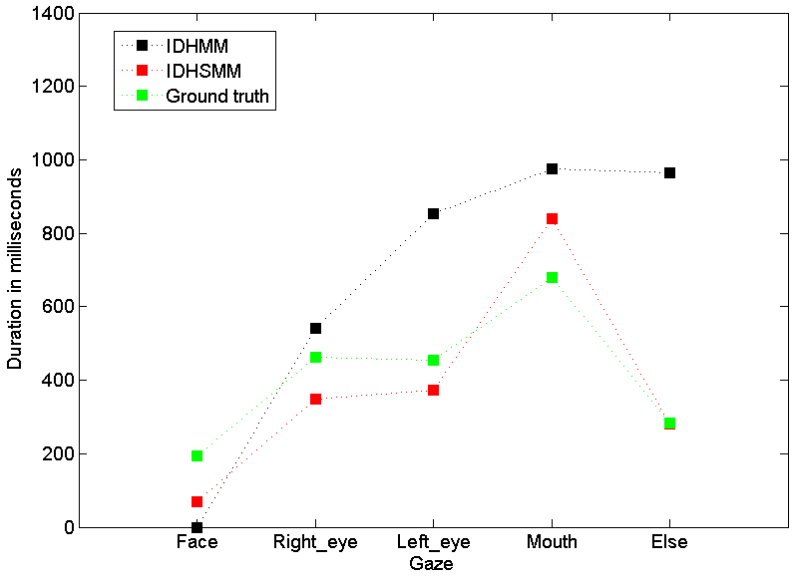


Figure 13: Comparison of mean duration of each region of interest between IDHMM and IDHSMM (for all subjects). Notice that no region of interest corresponding to "Face" was generated by IDHMM.

5 Conclusions

In this paper, we presented a comparative study of behavioral models designed to model face-to-face social interaction. A first model called IDHMM is introduced in which subtask-specific sensory-motor HHMs are trained and split into sensory HMMs for subtask recognition and motor HMMs for motor generation. Short-term Viterbi with a limited horizon is used to perform incremental recognition and generation. We have seen that even with low thresholds (up to one frame ahead), performances of the model are not significantly degraded. A remarkable property of this behavior model is the estimation of behavioral proximities between interlocutors. This could be exploited for

social evaluation but also to organize and select behavior models most suited to an unknown subject. Compared to classic classifiers (SVMs and decisions trees), the IDHMM model showed better performances thanks to its sequential modeling properties. In addition, classifiers like SVMs could result in good performance if a certain memory (~2 seconds in our case) was included in the input observations. Moreover, we show that data-driven dimensioning/initialization of the IDHMM improved the modeling of SMS cycles, and that explicit modeling of state duration improves substantially the generation figures.

In this paper, we argue that the sole quest for performance, ignoring properties of the generated signals (e.g. here state durations, loops patterns) may be misleading. One contribution of our paper is a thorough evaluation of the generated interactions with respect to properties that go beyond pure classification performance. The data mining, machine learning and objective evaluations have been facilitated by the repetitive structure of our interactions that maximize statistical coverage of IU realizations. Our next challenge is to scale up our behavioral models to larger sets of observations and IUs using more sparse interaction scores. Another challenge is to extract a minimal set of generic IUs that can bootstrap developmental learning of more complex interaction tasks and adapt to various users and situations thanks to principal modes of variation such as sketched in section 4.2.

Note also that the evaluation figures are totally objective and performed off-line. The statistically significant differences between performances do not prejudge for the results of on-line processing: small differences in the generated scores may result in large differences in reactive human behaviors and interaction management. For that reason, we are currently implementing those models on iCub robot, put the robot on a real face to-face interaction and get a subjective evaluation of the relevance of our models. We recently recorded also richer scenarios (with a larger number and higher branching factor of IUs) and larger sensory-motor scores (head and hand gestures in addition to gaze). We plan to apply our models to this new data, explore the best models, and then implement them on the robot.

6 Acknowledgements

This research is financed by the Rhône-Alpes ARC6 research council and the ANR-14-CE27-0014 SOMBRERO.

7 References

[1] A. Kendon, R. M. Harris, M. R. Key, et International Congress of Anthropological and Ethnological Sciences, *Organization of behavior in face-to-face interaction*. The Hague; Chicago: Mouton ; Distributed in the USA and Canada by Aldine, 1975.

[2] S. Scherer, S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, et L.-P. Morency, « Perception markup language: towards a standardized representation of perceived nonverbal behaviors », in *Intelligent Virtual Agents*, 2012, p. 455–463.

[3] J. L. Lakin, V. E. Jefferis, C. M. Cheng, et T. L. Chartrand, « The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry », *J. Nonverbal Behav.*, vol. 27, n° 3, p. 145‑162, sept. 2003.

[4] G. Bailly, « Boucles de perception-action et interaction face-à-face », *Rev. Franccaise Linguist. Appliquée*, vol. 13, n° 2, p. 121–131, 2009.

[5] G. Bailly, S. Raidt, et F. Elisei, « Gaze, conversational agents and face-to-face communication », *Speech Commun.*, vol. 52, n° 6, p. 598‑612, juin 2010.

[6] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, et M. Schroeder, « Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing », *IEEE Trans. Affect. Comput.*, vol. 3, n° 1, p. 69‑87, 2012.

[7] K. Otsuka, « Conversation Scene Analysis [Social Sciences] », *IEEE Signal Process. Mag.*, vol. 28, n° 4, p. 127‑131, 2011.

[8] D. Gatica-Perez, « Automatic nonverbal analysis of social interaction in small groups: A review », *Image Vis. Comput.*, vol. 27, n° 12, p. 1775–1787, 2009.

[9] A. Pentland, T. Choudhury, N. Eagle, et P. Singh, *Human dynamics: computation for organizations*. 2005.

[10] T. Choudhury et A. Pentland, « Characterizing Social Interactions Using the Sociometer », in *Proceedings of NAACOS 2004*, 2004.

[11] J. R. Curhan et A. Pentland, « Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes », *J. Appl. Psychol.*, vol. 92, n° 3, p. 802‑811, mai 2007.

[12] K. Otsuka, H. Sawada, et J. Yamato, « Automatic inference of cross-modal nonverbal interactions in multiparty conversations: "who responds to whom, when, and how?" from gaze, head gestures, and utterances », in *Proceedings of the 9th international conference on Multimodal interfaces*, New York, NY, USA, 2007, p. 255–262.

[13] D. Zhang, D. Gatica-Perez, S. Bengio, et I. McCowan, « Modeling individual and group actions in meetings with layered HMMs », *Multimed. IEEE Trans. On*, vol. 8, n° 3, p. 509–520, 2006.

[14] S. Petridis et M. Pantic, « Audiovisual discrimination between laughter and speech », in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, 2008, p. 5117‑5120.

[15] N. Fragopanagos et J. G. Taylor, « Emotion recognition in human–computer interaction », *Neural Netw.*, vol. 18, n° 4, p. 389‑405, mai 2005.

[16] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, et K. Karpouzis, « Modeling Naturalistic Affective States via Facial and Vocal Expressions

Recognition », in *Proceedings of the 8th International Conference on Multimodal Interfaces*, New York, NY, USA, 2006, p. 146–154.

[17] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, et S. Kollias, « Modeling Naturalistic Affective States Via Facial, Vocal, and Bodily Expressions Recognition », in *Artifical Intelligence for Human Computing*, T. S. Huang, A. Nijholt, M. Pantic, et A. Pentland, Éd. Springer Berlin Heidelberg, 2007, p. 91‑112.

[18] S. Banerjee et A. I. Rudnicky, « Using simple speech–based features to detect the state of a meeting and the roles of the meeting participants », 2004.

[19] D. B. Jayagopi, H. Hung, C. Yeo, et D. Gatica-Perez, « Modeling dominance in group conversations using nonverbal activity cues », *Audio Speech Lang. Process. IEEE Trans. On*, vol. 17, n° 3, p. 501–513, 2009.

[20] D. Gatica-Perez, « Analyzing group interactions in conversations: a review », in *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, 2006, p. 41–46.

[21] I. de Kok et D. Heylen, « Integrating Backchannel Prediction Models into Embodied Conversational Agents », in *Intelligent Virtual Agents*, Y. Nakano, M. Neff, A. Paiva, et M. Walker, Éd. Springer Berlin Heidelberg, 2012, p. 268‑274.

[22] M. K. Neff, « Gesture modeling and animation based on a probabilistic re-creation of speaker style. », *ACM Trans Graph*, vol. 27, 2008.

[23] H. Admoni et B. Scassellati, « Data-Driven Model of Nonverbal Behavior for Socially Assistive Human-Robot Interactions », in *Proceedings of the 16th International Conference on Multimodal Interaction*, New York, NY, USA, 2014, p. 196–199.

[24] S. P. Lee, J. B. Badler, et N. I. Badler, « Eyes Alive », in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, 2002, p. 637–644.

[25] L.-P. Morency, I. de Kok, et J. Gratch, « A probabilistic multimodal approach for predicting listener backchannels », *Auton. Agents Multi-Agent Syst.*, vol. 20, n° 1, p. 70‑84, janv. 2010.

[26] I. de Kok, D. Heylen, et L.-P. Morency, « Speaker-adaptive Multimodal Prediction Model for Listener Responses », in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, New York, NY, USA, 2013, p. 51–58.

[27] J. Lee et S. Marsella, « Modeling Speaker Behavior: A Comparison of Two Approaches », in *Intelligent Virtual Agents*, Y. Nakano, M. Neff, A. Paiva, et M. Walker, Éd. Springer Berlin Heidelberg, 2012, p. 161‑174.

[28] C.-M. Huang et B. Mutlu, « Learning-based Modeling of Multimodal Behaviors for Humanlike Robots », in *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, New York, NY, USA, 2014, p. 57–64.

[29] Y. Mohammad, T. Nishida, et S. Okada, « Unsupervised simultaneous learning of gestures, actions and their associations for Human-Robot Interaction », in *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009*, 2009, p. 2537‑2544.

[30] Y. Mohammad et T. Nishida, « Learning interaction protocols using Augmented Baysian Networks applied to guided navigation », in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, p. 4119‑4126.

[31] J. F. Ferreira, M. Castelo-Branco, et J. Dias, « A hierarchical Bayesian framework for multimodal active perception », *Adapt. Behav.*, vol. 20, n° 3, p. 172‑190, juin 2012.

[32] S. Levine, P. Krähenbühl, S. Thrun, et V. Koltun, « Gesture Controllers », in *ACM SIGGRAPH 2010 Papers*, New York, NY, USA, 2010, p. 124:1–124:11.

[33] K. R. Thórisson, « Natural Turn-Taking Needs No Manual: Computational Theory and Model, from Perception to Action », in *Multimodality in Language and Speech Systems*, B. Granström, D. House, et I. Karlsson, Éd. Springer Netherlands, 2002, p. 173‑207.

[34] C. E. Ford, « Contingency and Units in Interaction », *Discourse Stud.*, vol. 6, n° 1, p. 27‑52, janv. 2004.

[35] J. Lee, S. Marsella, D. Traum, J. Gratch, et B. Lance, « The Rickel Gaze Model: A Window on the Mind of a Virtual Human », in *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, Berlin, Heidelberg, 2007, p. 296–303.

[36] J. Rickel et W. L. Johnson, « Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control », *Appl. Artif. INTELLIGENCE*, vol. 13, p. 343–382, 1998.

[37] S. Marsella, J. Gratch, et J. Rickel, « Expressive Behaviors for Virtual Worlds », in *Life-Like Characters*, H. Prendinger et M. Ishizuka, Éd. Springer Berlin Heidelberg, 2004, p. 317‑360.

[38] L. R. Rabiner, « A tutorial on hidden markov models and selected applications in speech recognition », in *Proceedings of the IEEE*, 1989, p. 257–286.

[39] Y. Bengio et P. Frasconi, « Input-output HMMs for sequence processing », *IEEE Trans. Neural Netw.*, vol. 7, n° 5, p. 1231‑1249, sept. 1996.

[40] D. Gatica-Perez, « Analyzing group interactions in conversations: a review », in *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*, 2006, p. 41–46.

[41] R. Šrámek, B. Brejová, et T. Vinař, « On-line Viterbi Algorithm and Its Relationship to Random Walks », *arXiv:0704.0062*, mars 2007.

[42] J. Bloit et X. Rodet, « Short-time Viterbi for online HMM decoding: Evaluation on a real-time phone recognition task », in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, 2008, p. 2121‑2124.

[43] C. Y. Goh, J. Dauwels, N. Mitrovic, M. T. Asif, A. Oran, et P. Jaillet, « Online map-matching based on Hidden Markov model for real-time traffic sensing applications », in *2012 15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2012, p. 776‑781.

[44] S. Yu, « Hidden semi-Markov models », *Artif. Intell.*, 2010.

[45] J. D. Ferguson, « Variable Duration Models for Speech », *Symp Appl. Hidden Markov Models Text Speech Inst. Def. Anal. Princet. NJ*, p. 143‑179, oct. 1980.

[46] S. E. Levinson, « Continuously variable duration hidden Markov models for automatic speech recognition », *Comput. Speech Lang.*, vol. 1, n° 1, p. 29‑45, mars 1986.

[47] D. Kulp, D. Haussler, M. G. Reese, et F. H. Eeckman, « A generalized hidden Markov model for the recognition of human genes in DNA », *Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol.*, vol. 4, p. 134‑142, 1996.

[48] M. Russell, « A segmental HMM for speech pattern modelling », in *, 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993. ICASSP-93*, 1993, vol. 2, p. 499‑502 vol.2.

[49] P. Ramesh et J. G. Wilpon, « Modeling state durations in hidden Markov models for automatic speech recognition », in *, 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92*, 1992, vol. 1, p. 381‑384 vol.1.

[50] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, et T. Kitamura, « Hidden Semi-Markov Model Based Speech Synthesis », in *in Proc. of ICSLP, 2004*, 2004.

[51] W. P. Pierre Lanchantin, « UNSUPERVISED NON STATIONARY IMAGE SEGMENTATION USING TRIPLET MARKOV CHAINS », 2004.

[52] S. Hongeng et R. Nevatia, « Large-scale event detection using semi-hidden Markov models », in *Ninth IEEE International Conference on Computer Vision, 2003. Proceedings*, 2003, p. 1455‑1462 vol.2.

[53] K. Squire, « HMM-based semantic learning for a mobile robot, Ph.D. dissertation », Ph.D. dissertation, University of Illinois at Urbana-Champaign.

[54] S. Yu, « Multiple tracking based anomaly detection of mobile nodes », in *2005 2nd International Conference on Mobile Technology, Applications and Systems*, 2005, p. 5 pp.‑5.

[55] S. C. Schmidler, J. S. Liu, et D. L. Brutlag, « Bayesian segmentation of protein secondary structure », *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, vol. 7, n° 1‑2, p. 233‑248, avr. 2000.

[56] J. Bulla et I. Bulla, « Stylized facts of financial time series and hidden semi-Markov models », *Comput. Stat. Data Anal.*, vol. 51, n° 4, p. 2192‑2209, déc. 2006.

[57] C. Mitchell, M. Harper, L. Jamieson, et C. T. M, « On the Complexity of Explicit Duration HMMs », in *IEEE Transactions on Speech and Audio Processing*, 1995, p. 213–217.

[58] S. Yu et H. Kobayashi, « An efficient forward-backward algorithm for an explicit-duration hidden Markov model », *IEEE Signal Process. Lett.*, vol. 10, n° 1, p. 11‑14, janv. 2003.

[59] H. K. Shun-Zheng Yu, « Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden Markov model. », *IEEE Trans. Signal Process.*, vol. 54, p. 1947‑1951, 2006.

[60] S. Baron-Cohen, *Mind Reading: The Interactive Guide to Emotions*, Édition : Cdr. London u.a.: Jessica Kingsley Publishers, 2004.

[61] HTK, The Hidden Markov Model Toolkit, *http://htk.eng.cam.ac.uk/*. .

[62] M. Dunham et K. Murphy, *PMTK3: Probabilistic modeling toolkit for Matlab/Octave, http://code.google.com/p/pmtk3/*. .

[63] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten, « The WEKA data mining software: an update », *SIGKDD Explor Newsl*, vol. 11, n° 1, p. 10–18, nov. 2009.

[64] V. Levenshtein, « Binary Codes Capable of Correcting Deletions, Insertions and Reversals », *Sov. Phys. Dokl.*, vol. 10, n° 8, p. 707‑710, févr. 1966.

[65] A. Mihoub, G. Bailly, et C. Wolf, « Social Behavior Modeling Based on Incremental Discrete Hidden Markov Models », in *Human Behavior Understanding*, A. A. Salah, H. Hung, O. Aran, et H. Gunes, Éd. Springer International Publishing, 2013, p. 172‑183.

[66] A. Mihoub, G. Bailly, et C. Wolf, « Modeling Perception-Action Loops: Comparing Sequential Models with Frame-Based Classifiers », *ACM Hum. Agent Interact.*, 2014.

[67] D. C. Richardson, R. Dale, et K. Shockley, « Synchrony and swing in conversation: coordination, temporal dynamics, and communication », in *Embodied Communication in Humans and Machines*, I. Wachsmuth, M. Lenzen, et G. Knoblich, Éd. Oxford University Press, 2008, p. 75‑94.