Social behavior modeling based on Incremental Discrete Hidden Markov Models

Alaeddine Mihoub, Gérard Bailly, Christian Wolf

Social behavior modeling based on Incremental Discrete Hidden Markov Models

Alaeddine Mihoub[1, 2], Gérard Bailly[1], Christian Wolf[2]

[1] GIPSA-Lab, Speech & Cognition department, Grenoble, France
[2] LIRIS, Lyon, France

Abstract. Modeling multimodal face-to-face interaction is a crucial step in the process of building social robots or users-aware Embodied Conversational Agents (ECA). In this context, we present a novel approach for human behavior analysis and generation based on what we called "Incremental Discrete Hidden Markov Model" (IDHMM). Joint multimodal activities of interlocutors are first modeled by a set of DHMMs that are specific to supposed joint cognitive states of the interlocutors. Respecting a task-specific syntax, the IDHMM is then built from these DHMMs and split into i) a *recognition model* that will determine the most likely sequence of cognitive states given the multimodal activity of the interlocutor, and ii) a *generative model* that will compute the most likely activity of the speaker given this estimated sequence of cognitive states. Short-Term Viterbi (STV) decoding is used to incrementally recognize and generate behavior. The proposed model is applied to parallel speech and gaze data of interacting dyads.

Keywords: Face to face interaction, behavior model, action-perception loops, cognitive state recognition, gaze generation, HMMs, Online Viterbi decoding, latency.

1 Introduction

Face to face interaction is one of the most basic forms of communication for the human being in daily life [1]. Nevertheless, it remains a complex bi-directional multimodal phenomenon in which interlocutors continually convey, perceive and interpret the other person's verbal and nonverbal messages and signals [2]. Indeed, co-verbal cues [3] – such as body posture, arm/hand gestures (e.g. beat, deictic and iconic), head movement (e.g. node and tilt), facial expressions (e.g. frowning), eye gaze, eyebrow movement, blinks, as well as nose wrinkling and lips moistening – are largely involved in the decoding and encoding of linguistic and non-linguistic information. Several authors have notably claimed that these cues strongly participate in maintaining mutual attention and social glue [4][5].

Hence, social robots or conversational agents capable of ensuring a natural and multimodal communication should cope with complex perception-action loops that should mimic complex human behavior. In other terms, the social robot must be able to accomplish two main functionalities: (1) interaction analysis and (2) multimodal

behavior synthesis. In this context, we present a statistical modeling framework for capturing regularities of multimodal joint actions during face-to-face interaction, which allows us to achieve both interaction analysis and behavior synthesis. More precisely, this framework is based on the assumption that reactions to other's actions are ruled by the estimation of the underlying chaining of the cognitive states of the interlocutors.

The paper is organized as follows: The next section reviews state-of-the art of face to face interaction nonverbal analysis and then the behavior generation systems. Our modeling framework and its current implementation are introduced in section 3. Section 4 illustrates its modeling performance using speech and gaze data collected in a previous experiment [6] and shows the results. Finally, discussions and our conclusion are presented in section 5.

2 Related work

Face to face interaction analysis represents an emerging research area due to the increasing awareness of the scientific challenge and the diversity of applications. Actually automatic analysis treats many issues [7], among which can be mentioned: addressing, turn taking, activity recognition, roles, degree of interest or engagement, state of mind (e.g. neutral, curious, confused, amused) and dominance. A large number of models were proposed to cope with these problems. For instance, Otsuka et al [8] estimate turn taking ("who responds to whom and when?") with a Dynamic Bayesian Network consisting of three layers: (1) at the bottom, the behavior layer (contains head gestures and utterances); (2) in the middle, the interaction layer (contains gaze patterns); (3) at the top, the regime layer (contains conversations regimes). Only the first layer is observable, the others are latent and need to be estimated. To recognize group actions, Zhang et al [9] proposed a two layered HMM, where the first layer estimates individual actions from raw audio-visual data. The second one estimates group actions taking in consideration the results of the first layer. Conditional Random Fields are used in [10] for automatic role recognition in multiparty conversations. First, speaker diarization is applied to list turns; second, acoustic features are extracted from turns and finally, features vectors are mapped into a sequence of roles. More complete reviews on issues and models related to nonverbal analysis of social interaction can be found in [11][7].

In the context of multimodal behavior generation, several platforms have been proposed for humanoid robots and virtual agents. Cassel et al. [12] notably developed the BEAT ("Behavior Expression Animation Toolkit") system which allows from textual input to synthesize appropriate and synchronized behaviors with speech such as iconic gestures, eye gaze and intonation. The nonverbal behavior is assigned on the basis of linguistic and contextual analysis relying on a set of rules extracted from research on human conversational behavior. Krenn [13] introduced the NECA ("Net Environment for Embodied Emotional Conversational Agents") project which aims to develop a platform for the implementation of emotional conversational agents for Web-based applications. This system controls a complete scene generator and provides an ECA

with communicative (e.g eye brow raising, head nodes) as well as non communicative behavior (e.g physiological breathing, walking/moving from one location to another). Another major contribution of the NECA project is Gesticon [14] which consists of a repository of predefined co-verbal gestures and animations that can be accessed via functional descriptors. Gesticon is based on a general specification that may drive both physical and virtual agents. Another interesting system called "MAX", the "Multimodal Assembly eXpert", has been developed by Kopp [15]. The system allows interacting, in virtual reality environment, with a virtual agent and doing collaborative tasks. MAX is able to generate reactive and deliberative action using synthetic speech, gaze, facial expression, and gestures.

These different systems have many similarities: multimodal actions are selected, scheduled and combined according to rules that describe a sort of grammar of behaviors. The SAIBA framework [16] is an international effort to establish a unique platform and therefore speed up advancements in the field. It is organized into three main components: "Intent planning", "Behavior planning" and "Behavior realization". SAIBA adopted the Gesticon from the NECA platform and introduced two novel Markup Languages, the Behavior Markup Language (BML) [17] and the Functional Markup Language (FML) [18]. It is important to notice that SAIBA offers a general framework for building behavioral models. In fact, the processing within each component and its internal structure is treated as a "black box" and it is the researchers' responsibility to fill the boxes with their specific transducers. Through FML and BML, SAIBA aims at normalizing data types and information flows between different levels of representation of the behavior and bridge the gap between different modules: FML represents the output of the "Intent planning" component and BML the output of the "Behavior planning" one. Many systems have adopted the SAIBA framework, notably SmartBody [19] and the GRETA platform [20].

Human interactions are paced by multi-level perception-action loops [21] and one major missing aspect of the SAIBA was the perception dimension. The Perception Markup Language (PML) [2] has been recently introduced to fill this gap. It is the first step towards a standardized representation of perceived nonverbal behaviors. PML was inspired by the previous efforts in the field of non verbal behavior generation (FML and BML) and was designed in synergy with these standards. If PML has been equipped with the capability to carry uncertainty but the link between the uncertain perceptual representations and actions remains an open question. In the next section we will present our general behavior model which combines PML, FML and BML levels into a joint multimodal representation of task-specific human behavior. But unlike pre-mentioned rule-based models (BEAT, NECA, etc), this model relies on machine learning to organize sequences of percepts and actions into so-called joint behavioral states using Hidden Markov Models (HMMs).

3 General behavior Model

This section presents a probabilistic/statistical approach for designing a dynamic model for the generation of pertinent multimodal behavior for a humanoid robot or an

ECA engaged in a collaborative task with a human partner. This model should thus be able to perceive and understand the partner's actions on their joint environment and generate adequate actions that should reflect its current understanding of the evolution of the joint plan.

A complete interaction can be seen as a sequence of discrete tasks, sub-tasks or activities [11]. In the following, we will consider a situated conversation as a sequence of cognitive states that structure the joint behaviors of the conversation partners. In our model, we dispose of *P* cognitive states; each cognitive state is modeled by a single Discrete Hidden Markov Model ($\lambda_p = (A_p, B_p, \Pi_p)_{p=1..P}$) whose *n* hidden states model the co-variations of the partners' behaviors. The proper chaining of these HMMs obeys to a task-specific syntax and results from lawful mutual attention and collaborative actions. Hence, the whole interaction is modeled by a global Discrete HMM ($\lambda = (A, B, \Pi)$) that concatenates the different single models. Thus the global DHMM λ is composed of *N* hidden states (*N*=*nP*). As a matter of fact, the selection and sequencing of these HMMs is equivalent to the ordering of instructions in the FML level within the SAIBA framework. Consequently, the problem of 'intent planning' is solved by the process of HMM states decoding [22], usually performed by the Viterbi algorithm.

As mentioned before, HMM states are associated with homogenous joint sensory-motor behaviors: the observation vector $O = (o_t)_{t=1..T}$ is in fact composed of two streams: (1) the sensory stream (o_t^p) collects perceptual cues and roughly correspond to the low-level PML level in the SAIBA framework; (2) the motor stream (o_t^a) is responsible for initiating actions and roughly corresponds to the BML level in the SAIBA framework. The observation vector is then defined as follows:

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

Note that the sensory stream may include sensory consequences of actions. These may be of different natures: efferent copies of actions, accompanying proprioceptive or exteroceptive signals. Compared to the Gesticon, our sensory-motor states (**Fig. 1**) intrinsically associate actions and percepts and do not differentiate between the perceptual responses of an action and motor responses for a perceived event that are appropriate to the current joint cognitive state.

3.1 Training, sensory-motor state alignment, cognitive state recognition, and action generation

The training process is as follow: Each individual model is trained separately; then from single HMMs we get local emission Matrices $(B_p)_{p=1..P}$ and simply concatenate them to build the global emission matrix *B*. Like-wise, 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 are emitted according to these input cues. For that reason, once we get the global trained HMM, two models are extracted: a recognition model λ_R and a generative model λ_G

with a modified structure for the emission matrix *B*. For λ_R only perception observations are selected (i.e. $o_t = o_t^p$) and for λ_G only action observations are selected (i.e. $o_t = o_t^a$). The perception for action loop combines recognition and synthesis: λ_R decodes percepts and performs the sensory-motor states alignment while λ_G further generates the adequate actions.

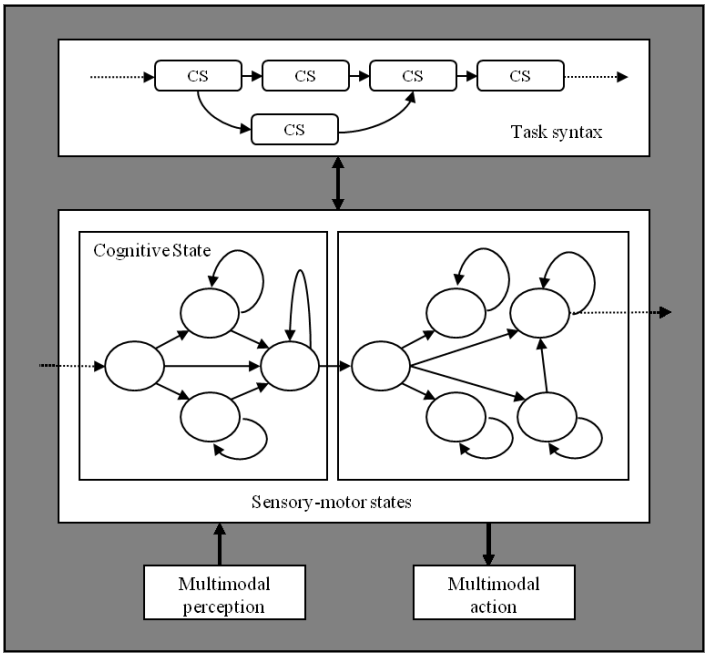

Fig. 1. Management of perception-action loops in a probabilistic scheme linking observation, states and task syntax (sequence of cognitive states)

3.2 Incremental Discrete Hidden Markov Model

The Viterbi algorithm allows estimating of the most likely state sequence S^* according to an observed sensory stream O and a HMM model λ:

$$S^* = arg\max_S(S|O,\lambda) \tag{2}$$

This alignment between observations and states is usually performed in two steps:

1. A forward step computes the partial likelihoods δ_t and stores the best predecessor for each state at each time frame in a matrix of backtracking pointers ψ_t.
2. A backtracking step on ψ_t builds the optimal path from the end of the observation sequence.

In order to exploit partial backtracking for on-line decoding, several solutions have been proposed that use a fixed sliding or overlapping window [23] [24] [25] [26]. It

consists of dividing the sequence into fixed-size inputs and then decodes them independently. An alternative approach consists of using an expending window and comparing partial paths until convergence to the same trajectory [27] [28] [29]. The central idea of the Short-Time Viterbi (STV) algorithm [28] and its variants is that the window is continuously expending forward until a convergence/fusion point is found. When this is the case, it shrinks from behind. 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. The BSTV algorithm is described briefly as follow:

```
initiate δ_1; ψ_1; a=1;
for each new frame b
  for each state j=1:N
      calculate δ_b(j) and ψ_b(j);
      backtracking: S*_{t-1} = ψ_t(j) with t=b:a+1;
      save the local path;
  end
  given all local paths find fusion point f;
  if (b-a<threshold and f exists)
      local path for t=a:f is selected; a=f+1;
  else if (b-a>=threshold)
      path with max likelihood is selected;
      f=b;a=b;
      b=b+1;
  else b=b+1;
end
```

Although, the optimal solution is not always selected, the latency is fully controlled. We will show that short latencies obtained in practice do not degrade significantly the performance of the decoder.

In the next section, we apply this Incremental Discrete Hidden Markov Model (IDHMM) to multimodal experimental speech and gaze data of computer-mediated dyadic conversations.

4 Experimental results

We used the dataset of Bailly et al. [6], 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 **Fig. 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 read and utter a sentence that the other sub-

ject (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 respondent to be highly attentive to the audiovisual signal.

The experiment was designed to study adaptation: one female speaker HL interacted with ten subjects (6 female colleagues, 3 female students and one male student), both as an instructor for ten sentences and as a respondent for another ten sentences.

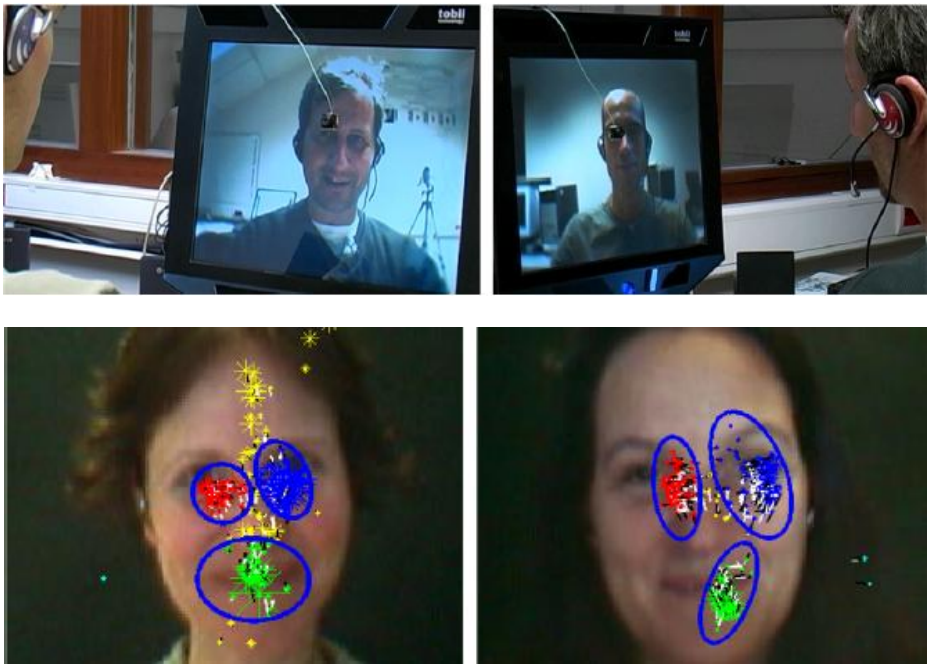

Fig. 2. Mediated face-to-face conversation [6]. Top: People sit in two different rooms and dialog through couples of cameras, screens, microphones and loudspeakers. Gaze of both interlocutors are monitored by two eye-trackers embedded in the TFT screens. Note that pinhole cameras and seats are positioned 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).

4.1 Data

The observation sensory streams consist here of discrete observations: the voice activity (cardinality 2: on/off) and ROI (cardinality 5) of the two speakers. The cognitive states (CS) have been labelled semi-automatically and corrected by hand. We distinguish between seven CS: reading, preparing to speak, speaking , waiting, listening, thinking and else (laughing, etc). These CS may occur for each speaker in three different roles: initiator, respondent or none (free interaction before, after and when exchanging roles).

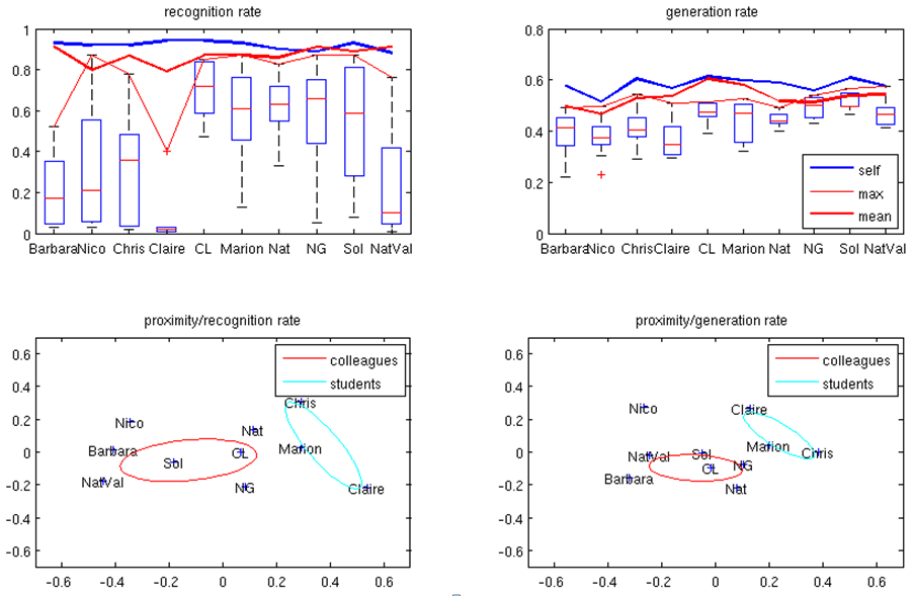

Fig. 3. Recognition (left) and generation (right) performance. Top: the performances of interlocutor-independent (II: dark red), interlocutor-dependent (ID: boxplots and maximum with light red) and self DHMMs (dark blue) are displayed for each interlocutor. Bottom, a MDS projection of the performances of the ID models cue proximities between interlocutor-specific behaviors: note its coherence with the a priori clustering of their social relations with HL.

4.2 Behavioral models

We tested the ability of DHMMs and IDHMM to estimate the cognitive state of the main subject "LN" given her voice activity (*v1*), gaze (*g2*) and voice activity of her conversational partner (*v2*), and predict her gaze behavior (*g1*). Consequently, we use the recognition model λ_R to decode $o_t = (v1, v2, g2)$ and next λ_G to generate the gaze (*g1*).

4.3 Results using DHMMs

We build and test different models in an offline mode using HTK [30]: for interlocutor-dependent (ID) vs. interlocutor-independent (II) models. For each interlocutor, the corresponding II model is trained on the other 9 interactions. Results are illustrated in **Fig. 3**: the mean recognition and correct generation rate of II models are respectively 93% and 56% (compared to a random assignment at 23% taking into account a priori distributions of ROI).

A multidimensional scaling analysis based on Kruskal's normalised STRESS1 criterion was performed on ID cognitive state recognition and gaze prediction errors (see bottom of **Fig. 3**). This analysis of proximity of behaviors nicely mirrors the a priori social relationships between HL and her interlocutors. 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.

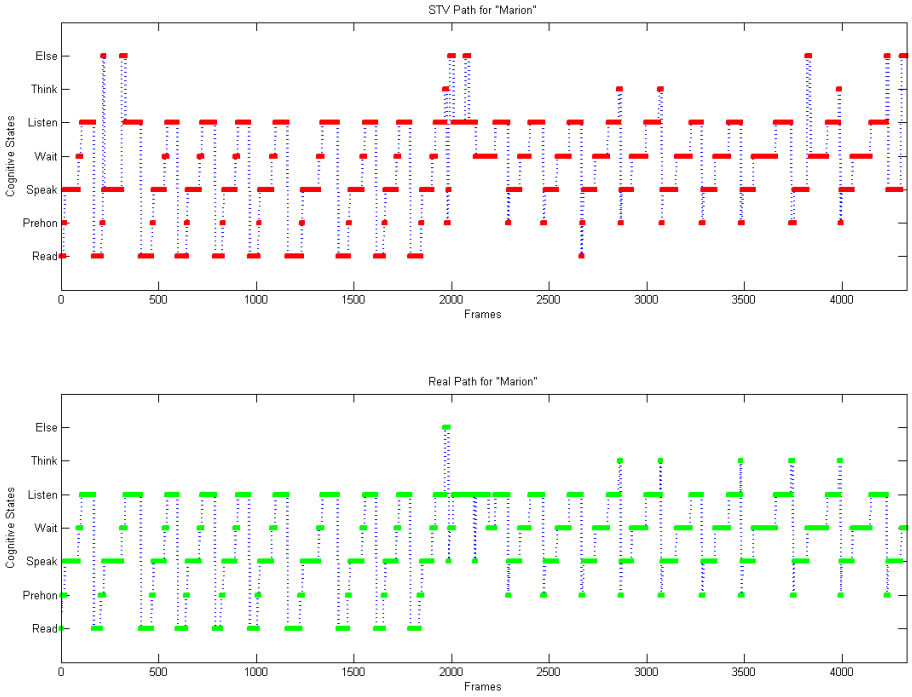

Fig. 4. Recognition path (for a specific interlocutor "Marion") using the incremental model (top) vs. ground truth (bottom).

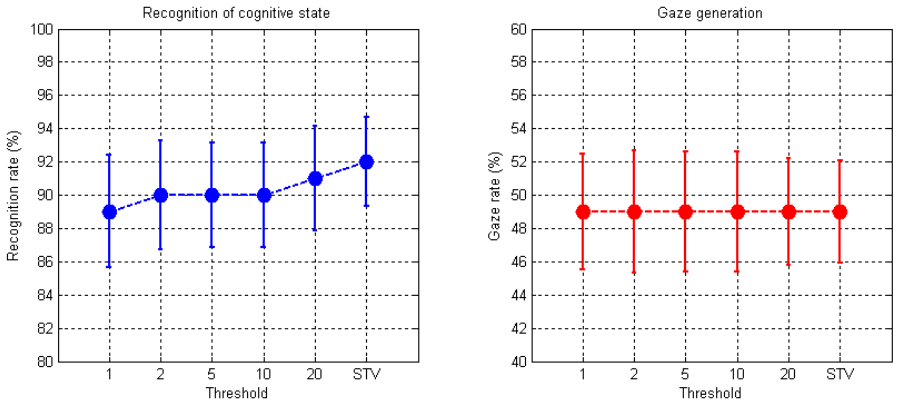

Fig. 5. Recognition and generation results using the incremental model

4.4 Results using IDHMM

HMMs are trained with HTK, then the BSTV algorithm and the global HMM are implemented in Matlab using PMTK3 toolkit [31]. The mean recognition rate of 92% shows that STV is able to capture the structure of the interaction (see **Fig. 4** and **Fig. 5**). It confirms also that STV performance is as good as an offline processing. However, the problem with STV is mastering the output delay. We observe that ~80% of latencies are fewer than 5 frames. However, maximum values could be very important. In our case, for the all subjects, the maximum latency was 259 frames which represent an unsuitable delay for real-time application. BSTV is used to control these delays. Theoretically, an optimal trade-off ought to be sought because of the inverse relationship between performance and latency. In our case, results (**Fig. 5**) have shown that our IDHMM is able to estimate the Viterbi path with low thresholds/latencies as well as for a long term processing (e.g. 90% for a threshold equal to 2). Moreover the mean generation performance (49%) is not affected and remains practically the same at all thresholds. While the full connectivity of the state transition matrix explains why almost 80% of latencies are fewer than 5 frames (i.e. deviations of the local path to the global path may be rapidly reconnect when robust cues are encountered), another important factor is the syntax of the task: the chaining of subtasks is very regular and highly constraints the alignment of cognitive states.

5 Conclusions

We have proposed a modeling framework for the recognition and the generation of joint multimodal behavior. Sub-task sensory-motor HHM are trained and split into sensory HMM for sub-task recognition and motor HMM for motor generation. Short-term Viterbi with a limited horizon is used to perform incremental recognition and generation. We showed that even with low thresholds, performances of the model were not significantly degraded. This first model will be extended to the joint modeling of discrete and continuous observations, notably taking into account the strengths of trajectory HMM.

A noteworthy property of these statistical behavior models is the estimation of behavioral proximities/distances between subjects. This could be exploited for social evaluation but also to organize and select behavior models most adapted to an unknown interlocutor.

Due to lack of space, many technical details such as the initialization and training of Markov models for discrete observations and fully-connected states deserve in-depth analysis and require more research effort. In particular, performance would largely benefit from the modeling of state durations (here related to gaze fixations).

Acknowledgments

This research is financed by the Rhône-Alpes ARC6 research council.

References

[1] K. Otsuka, "Multimodal Conversation Scene Analysis for Understanding People's Communicative Behaviors in Face-to-Face Meetings," pp. 171–179, 2011.

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

[3] M. Argyle, *Bodily Communication*. Taylor & Francis, 1975.

[4] J. L. Lakin, V. E. Jefferis, C. M. Cheng, and T. L. Chartrand, "The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry," *Journal of Nonverbal Behavior*, vol. 27, no. 3, pp. 145–162, Sep. 2003.

[5] S. Kopp, "Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors," *Speech Commun.*, vol. 52, no. 6, pp. 587–597, juin 2010.

[6] G. Bailly, S. Raidt, and F. Elisei, "Gaze, conversational agents and face-to-face communication," *Speech Communication*, vol. 52, no. 6, pp. 598–612, juin 2010.

[7] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[8] K. Otsuka, H. Sawada, and 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, pp. 255–262.

[9] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered HMMs," *Multimedia, IEEE Transactions on*, vol. 8, no. 3, pp. 509–520, 2006.

[10] H. Salamin and A. Vinciarelli, "Automatic Role Recognition in Multiparty Conversations: An Approach Based on Turn Organization, Prosody, and Conditional Random Fields," *IEEE Transactions on Multimedia*, vol. 14, no. 2, pp. 338–345, 2012.

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

[12] J. Cassell, H. Vilhjalmsson, and T. Bickmore, *BEAT: the Behavior Expression Animation Toolkit*. 2001.

[13] B. Krenn, "The NECA project: Net environments for embodied emotional conversational agents," in *Proc. of Workshop on emotionally rich virtual worlds with emotion synthesis at the 8th International Conference on 3D Web Technology (Web3D), St. Malo, France*, 2003, vol. 35.

[14] B. Krenn and H. Pirker, "Defining the gesticon: Language and gesture coordination for interacting embodied agents," in *Proc. of the AISB-2004 Symposium on Language, Speech and Gesture for Expressive Characters*, 2004, pp. 107–115.

[15] S. Kopp, B. Jung, N. Lessmann, and I. Wachsmuth, "Max - A Multimodal Assistant in Virtual Reality Construction," *KI*, vol. 17, no. 4, p. 11, 2003.

[16] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson, "Towards a Common Framework for Multimodal Generation: The Behavior Markup Language," in *INTERNATIONAL CONFERENCE ON INTELLIGENT VIRTUAL AGENTS*, 2006, pp. 21–23.

[17] H. Vilhjálmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. Marshall, and C. Pelachaud, "The behavior markup language: Recent developments and challenges," in *Intelligent virtual agents*, 2007, pp. 99–111.

[18] D. Heylen, S. Kopp, S. C. Marsella, C. Pelachaud, and H. Vilhjálmsson, "The Next Step towards a Function Markup Language," in *Proceedings of the 8th international conference on Intelligent Virtual Agents*, Berlin, Heidelberg, 2008, pp. 270–280.

[19] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann, "Smartbody: Behavior realization for embodied conversational agents," in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, 2008, pp. 151–158.

[20] Q. A. Le and C. Pelachaud, "Generating Co-speech Gestures for the Humanoid Robot NAO through BML," in *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, E. Efthimiou, G. Kouroupetroglou, and S.-E. Fotinea, Eds. Springer Berlin Heidelberg, 2012, pp. 228–237.

[21] G. Bailly, "Boucles de perception-action et interaction face-à-face," *Revue fran\ccaise de linguistique appliquée*, vol. 13, no. 2, pp. 121–131, 2009.

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

[23] A. Seward, *Low-Latency Incremental Speech Transcription in the Synface Project*. .

[24] M. Ryynänen and A. Klapuri, "Automatic Bass Line Transcription from Streaming Polyphonic Audio," in *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. 1437–1440.

[25] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Map-matching for low-sampling-rate GPS trajectories," in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, New York, NY, USA, 2009, pp. 352–361.

[26] J. Yuan, Y. Zheng, C. Zhang, X. Xie, and G.-Z. Sun, "An Interactive-Voting Based Map Matching Algorithm," 2010, pp. 43–52.

[27] R. Šrámek, B. Brejová, and T. Vinař, "On-line Viterbi Algorithm and Its Relationship to Random Walks," *arXiv:0704.0062*, Mar. 2007.

[28] J. Bloit and 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, pp. 2121–2124.

[29] C. Y. Goh, J. Dauwels, N. Mitrovic, M. T. Asif, A. Oran, and 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, pp. 776–781.

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

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