Learning joint multimodal behaviors for face-to-face interaction: performance properties of statistical models

Gérard Bailly, Alaeddine Mihoub, Christian Wolf, Frédéric Elisei

Learning joint multimodal behaviors for face-to-face interaction: performance & properties of statistical models

Gérard Bailly[(1)] Alaeddine Mihoub[(1,2)] Christian Wolf[(2)] Frédéric Elisei[(1)]

[(1)] GIPSA-Lab, Université Grenoble Alpes/CNRS
11, rue des Mathématiques, St Martin d'Hères, France
firstname.lastname@gipsa-lab.fr

[(2)] Université de Lyon/CNRS
INSA-Lyon, LIRIS, UMR5205 F-69621, France
christian.wolf@liris.cnrs.fr

ABSTRACT

We evaluate here the ability of statistical models, namely Hidden Markov Models (HMMs) and Dynamic Bayesian Networks (DBNs), in capturing the interplay and coordination between multimodal behaviors of two individuals involved in a face-to-face interaction. We structure the intricate sensory-motor coupling of the joint multimodal scores by segmenting the whole interaction into so-called interaction units (IU). We show that the proposed statistical models are able to capture the natural dynamics of the interaction and that DBNs are particularly suitable for reproducing original distributions of so-called coordination histograms.

Theme: behavior coordination between animals, humans and robots

Keywords: human-human interaction; multimodal interaction; behavioral models; machine learning; statistical modeling.

1 INTRODUCTION

Deictic expressions [1] such as the famous "put that there" explored by Bolt [2] implies a tight coordination between gaze, head/torso/arm and finger pointing and speech [3, 4]. This context-dependent intermodal coordination is further affected by interleaving multimodal cues provided by the recipient of the information: responsive gaze cues, mimics as well as acoustic backchannels pace the effective encoding and decoding of the intended message. Several seminal works such as those of Richardson et al on swinging [5], MacFarland on respiration [6] and Bailly et al on gaze [7] have shown that the patterns of multimodal coordination are sensitive to cognitive requirements.

Coordinated behaviors are traditionally described with an *epistemic* approach: context-sensitive rules describe how speech events and gestural strokes respond and align with others' speech events and gestural strokes within discourse units. Initially developed for conversational agents, numerous rule-based systems have been proposed to cope with such a complex orchestration of multimodal streams: see notably the Ymir model proposed by Thórisson [8] or the series of mark-up languages developed more recently for handling multimodal communication and interaction [9-11].

More recently, *epigenetic* approaches have been proposed to learn patterns of coordination from sensory-motor experience. These data-driven approaches automatically infer the behavioral models from embodied and grounded interaction thanks to machine learning techniques and various learning strategies (observation, demonstration,etc). We introduce below two of these techniques.

2 STATISTICAL MODELS

Statistical models have been used for years for scene analysis, i.e. inferring semantic information from signals: speech and speaker recognition, visual scene analysis as well as inferring human activity, emotions or social features of the conversional partners. More recently, statistical models have been used to cope with behavior generation, e.g. see [12] for speech synthesis and [13] for gesture synthesis.

Multi-space probability distribution models [14] and various statistical models capturing the cross-correlations between time series have been further used to generate signals from observed ones. Otsuka et al. [15] proposed a Dynamic Bayesian Network (DBN) to estimate addressing and turn taking ("who responds to whom and when?") from speech and head gestures and generate appropriate gaze patterns. Morency et al [16] showed how 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). For more insights see also [17, 18]. More generally, these approaches use probabilistic graphical models because they provide a flexible and trainable representation of the dynamics of human behavior and capture spatiotemporal relationships between multimodal observations under uncertainty.

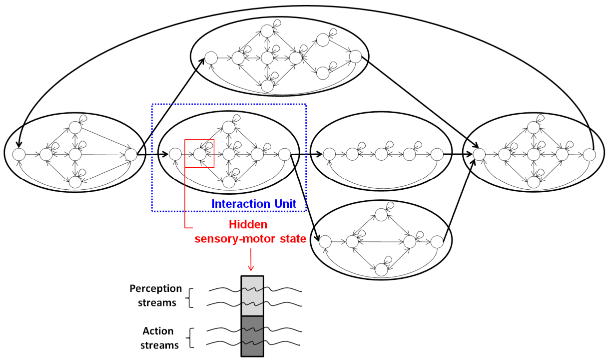

Figure 1. Mapping perception to action using HMM. Hidden sensory-motor states are trained using parallel perception and action streams. The structuring of these hidden states into *interaction units* facilitates the alignment of states and observations at training stage. At run time, states are inferred on the sole basis of the observed perception stream and generate the corresponding motor stream (see text for further explanation).

2.1 Hidden Semi-Markov Models (HSMMs)

Sensory-motor scores (i.e. perceptual input and motor output trajectories) are here supposed to be generated by unobserved (*hidden*) states of a Markov process. In a hierarchical model, these hidden sensory-motor states are organized into *interaction units* (IUs) [19] (such as thinking, informing, listening, taking turn, glazing over, etc). Each IU is thus responsible for modeling several parts of the sensory-motor scores that share common semantic or pragmatic information. The number, extend and 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 interaction units. A given IU can be seen as an instance of the *joint cognitive states* of the interacting dyads. A similar concept was used in [20] in which the authors propose a gaze model driven by the cognitive operations of a virtual agent. Note that disjoined or desynchronized cognitive states can be modeled via coupled or product HMMs.

Once trained on parallel perception and action streams (see Figure 1), sensory-motor HMMs are easily split into a *recognition* model that infers the optimal state sequence on the sole basis of the observed perception stream and a *synthesis* model that generates the most likely motor observations given the decoded sequence of states (see [21] for further explanation).

Incremental Hidden Semi-Markov Models (HSMMs) further combine the possibility to model state duration (also called residence time) and to infer states and generate actions with limited look ahead. The bounded version of the Short-Time Viterbi (STV) algorithm [22] slightly impairs the estimation of the IUs but with no substantial degradation of the generated actions [23].

2.2 Dynamic Bayesian Networks (DBNs)

Classical HMMs are characterized by a fixed dependency graph modeling the conditional independence relationships between random variables. Extensions to the standard HMM formalism have been proposed to cope with direct input/output dependencies [24]. By representing observations and hidden states as random variables, DBNs are particularly suitable for modeling the dynamics of multimodal behaviors in face-to-face interactions [25] [25], as their graphical structure can be arbitrary. These dependency structures between variables can be provided by an expert but numerous methods have been introduced to learn the network's structure automatically from data both for intra- and inter-frame conditional dependencies.

An example of such a dependency graph obtained for the data described below is displayed Figure 3. We nicely recover the causal relations between the gaze, pointing gesture, speech of the instructor and the final grasping gesture of the manipulator.

Figure 2. The "put-that-there" game.

3 MULTIMODAL INTERACTIVE DATA

We trained our statistical models on interactive data collected during a "put-that-there" game (see Figure 2). Models are trained to capture the behavior of the instructor. The perceptual and action streams are gathered by respectively analyzing the visual field of the instructor via a head-mounted camera and monitoring his speech, head, gaze and hand movements (see the deictic gesture in Figure 2) by motion capture devices. The scenario consists in a repetitive task that samples the working space: the instructor asks the manipulator to reproduce various cube arrangements on a chessboard according to a layout he is the only one to know. The task of the statistical model is to predict the gaze (FX) and hand movements (GT) of the instructor given his speech (SP) and the perceived gestures (MP) of the manipulator. All observations are discretized, e.g. alternative points of interest for gaze fixations (cubes, locations, hands, and manipulator's face), gesture strokes (grasp, pick-up, transport, and release), speech (cube name, source and target locations), etc. We distinguish between 6 interaction units: get instruction, seek cube, point to source location, indicate target location, check manipulation, and validate manipulation. Note that HMM sensory-motor states are fully connected: seeks, mutual attentions, errors and repetitions often result in state looping and acyclic graphs that may solicit several times the same hidden state within one interaction unit.

4 COORDINATION HISTOGRAMS

Observations are here discrete: modal events are thus generated each time the statistical model observes or generates a transient between successive observations. A coordination histogram (CH) for a couple of modalities is computed as follows: for each event, we search for the nearest event of the other modality and record the time lag between these two events. The CH for ground truth fixations with reference to input streams are compared in Figure 4 with CH for data generated by various statistical models. While DBN significantly outperforms HSMM – which significantly outperforms HMM – in terms of prediction performance [26], coordination patterns are also significantly closer to original ground truth data.

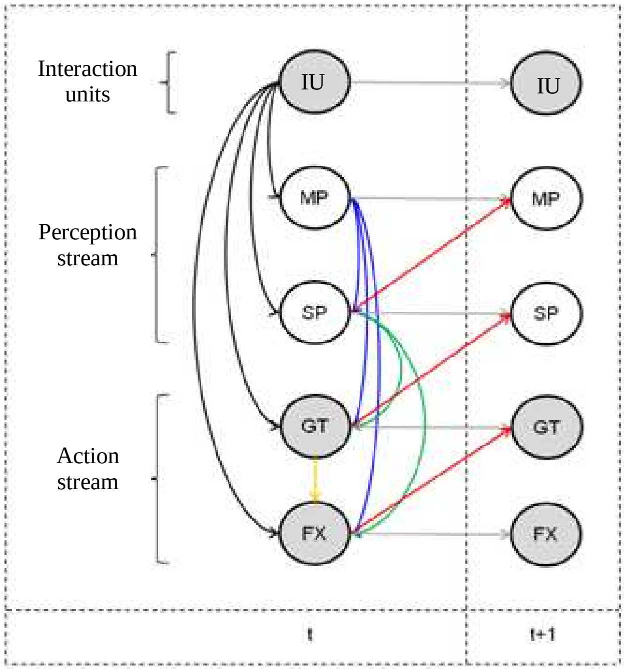

Figure 3: Dependency graph between interaction units, perception and action streams automatically inferred from "put-that-there" data for the DBN. Arrows of different colors (distinguishing intra- vs. inter-modality, intra- vs. inter-frame dependencies) cue significant dependencies between units and observations. Contrary to HMM, no latent states have been added here.

CONCLUSIONS

We present statistical multimodal behavioral models trained on multimodal data collected on dyads involved in a situated collaborative face-to-face interaction. We compared the performance of Hidden Markov Model (HMM), Hidden Semi-Markov Model (HSMM) and Dynamic Bayesian Network (DBN). We introduce the concept of coordination histogram that characterizes how different modalities synchronize between each other. DBN leads to the best performances in both interactive units recognition and behavior generation. It also displays a faithful coordination between generated trajectories compared to the ground truth. We suggest incorporating such behavioral characteristics for model assessment and coordination studies.

We plan to further implement the models on social robots in order to gather subjective evaluations and performative assessments.

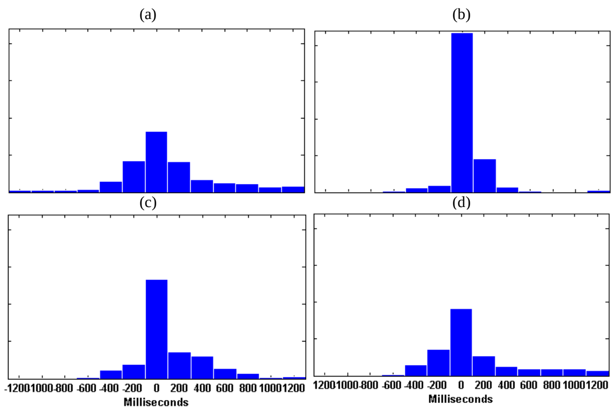

Figure 4: Coordination histograms of the instructor's gaze with the input streams (his own speech onsets and the onsets of the manipulator's strokes) for the ground truth (a) and the predicted gaze by HMM (b), HSMM (c) and DBN (d).

ACKNOWLEDGMENTS

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

5 REFERENCES

[1] Kranstedt, A., P. Kühnlein, and I. Wachsmuth, *Deixis in Multimodal Human Computer Interaction: An Interdisciplinary Approach*, in *Gesture-Based Communication in Human-Computer Interaction*, A. Camurri and G. Volpe, Editors. 2004, Springer Berlin, Heidelberg. p. 112-123.

[2] Bolt, R.A., *"Put-that-there": Voice and gesture at the graphics interface*. ACM SIGGRAPH Computer Graphics, 1980. 14(3): p. 262-270

[3] Rochet-Capellan, A., et al. *Does the number of syllables affect the finger pointing movement in a pointing-naming task ?* in *International Seminar on Speech Production*. 2008. Strasbourg - France. p. 257-260.

[4] Roustan, B. and M. Dohen. *Gesture and speech coordination : The influence of the relationship between manual gesture and speech*. in *Interspeech*. 2010. Makuhari, Japan. p. 498-501.

[5] Richardson, D.C., R. Dale, and K. Shockley, *Synchrony and swing in conversation: coordination, temporal dynamics, and communication*, in *Embodied Communication*, I. Wachsmuth, M. Lenzen, and G. Knoblich, Editors. 2008, Oxford University Press: Oxford, UK. p. 75-93.

[6] McFarland, D.H., *Respiratory markers of conversational interaction*. Journal of Speech and Hearing Research, 2001. 44: p. 128-143.

[7] Bailly, G., S. Raidt, and F. Elisei, *Gaze, conversational agents and face-to-face communication*. Speech Communication - special issue on Speech and Face-to-Face Communication, 2010. 52(3): p. 598–612.

[8] Thórisson, K., *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, and I. Karlsson, Editors. 2002, Kluwer Academic: Dordrecht, The Netherlands. p. 173–207.

[9] Vilhjalmsson, H., et al. *The Behavior Markup Language: recent developments and challenges*. in *International Conference on Intelligent Virtual Agents*. 2007. Paris: LNCS 4722. p. 99-111.

[10] Heylen, D., et al. *The next step towards a functional markup language*. in *Intelligent Virtual Agents (IVA)*. 2008. Tokyo. p. 37-44.

[11] Scherer, S., et al. *Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors*. in *International Conference on Intelligent Virtual Agents (IVA)*. 2012. Santa Cruz, CA.

[12] Zen, H., K. Tokuda, and A.W. Black, *Statistical parametric speech synthesis*. Speech Communication, 2009. 51: p. 1039–1064.

[13] Calinon, S., et al., *Learning and reproduction of gestures by imitation: An approach based on Hidden Markov Model and Gaussian Mixture Regression*. IEEE Robotics and Automation Magazine, 2010. 17(2): p. 44-54.

[14] Tokuda, K., et al., *Multi-space probability distribution HMM*. IEICE Transaction of Information and System, 2002. E85-D(3): p. 455-464.

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

[16] Morency, L.-P., I. de Kok, and J. Gratch, *A probabilistic multimodal approach for predicting listener backchannels*. Autonomous Agents and Multi-Agent Systems (AAMAS), 2010. 20: p. 70-84.

[17] Mohammad, Y., T. Nishida, and 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 (IROS)*. 2009. St. Louis, MO. p. 2537-2544.

[18] Ferreira, J.F., et al., *A Bayesian framework for active artificial perception*. IEEE Transactions on Cybernetics, 2013. 43(2): p. 699-711.

[19] Ford, C.E., *Contingency and units in interaction*. Discourse Studies, 2004. 6(1): p. 27-52.

[20] Lee, J., et al. *The Rickel gaze model: A window on the mind of a virtual human*. in *Intelligent Virtual Agents Conference*. 2007. Paris, France.

[21] Mihoub, A., G. Bailly, and C. Wolf. *Modelling perception-action loops: comparing sequential models with frame-based classifiers*. in *Human-Agent Interaction (HAI)*. 2014. Tsukuba, Japan. p. 309-314.

[22] Bloit, J. 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 (ICASSP)*. 2008. Las Vegas, NE. p. 2121-2124.

[23] Mihoub, A., G. Bailly, and C. Wolf, *Learning multimodal behavioral models for face-to-face social interaction*. Journal on Multimodal User Interfaces (JMUI), in revision.

[24] Bengio, Y. and P. Frasconi, *Input-output HMMs for sequence processing*. IEEE Transactions on Neural Networks, 1996. 7(5): p. 1231-1249.

[25] Huang, C.-M. and B. Mutlu. *Learning-based modeling of multimodal behaviors for humanlike robots*. in *ACM/IEEE international conference on Human-Robot Interaction (HRI)*. 2014. Bielefeld, Germany. p. 57-64.

[26] Mihoub, A., et al., *Graphical models for social behavior modeling in face-to face interaction*. Pattern Recognition Letters, submitted.