Speech in action: designing challenges that require incremental processing of self and others' speech and performative gestures

Gérard Bailly and Frédéric Elisei
firstname.familyname@gipsa-lab.fr
GIPSA-Lab, Univ. Grenoble-Alps & CNRS
Grenoble, France

Figure 1: Sample situations combining speech and deictic gestures while monitoring attendance's attention. Left: face-to-face "Put That There" task [11]. Right: robot instructing two students how to re-mount a jigsaw tool [2].

ABSTRACT

We advocate here challenges that elicit more intimate coordination between verbal communication and performative gestures.

CCS CONCEPTS

• **Computer systems organization** → Robotics.

KEYWORDS

speech, deictic gestures, gaze, incremental processing, HRI, challenge

1 INTRODUCTION

Several initiatives have already been conducted to shorten the bridge between robotics and speech technology: "HRI Face-to-Face: Gaze and speech communication" workshop at HRI 2013, "Speech & HRI" session at Interspeech 2017, VIHAR workshops (Vocal interactivity in-and-between Humans, Animals and Robots) 2016 [12] & 2019 [21], SLIVAR Dagstuhl seminar in 2020 [1]…Despite that computer science, automation and signal processing have other intersecting hot spots (HRI, IROS, etc), the research communities are still observing each other at distance, using off-the-shelves innovations from both camps whenever available.

2 SPEECH TECHNOLOGY AND ROBOTICS

Speech recognition and synthesis are mainly used as front-ends and back-ends of dialog systems, replacing textual i/o. In short, speech and gesture are often viewed as commands that the robot can interpret and respond to [9, 24]. As rapidly as possible, multimodal signals are converted into words, semantic frames, intentions, emotional tags …and vice-versa in order to update the state space of the artificial agent, embodied by a voice, a virtual agent or a physical robot. Just like other automatic processing of sensory information, the performance of current speech recognition is often strongly impaired in HRI due to wild environmental conditions [16], sensor motion [15] and ego-noise, etc (see section on "challenges on technologies and sensors for HRI" in the review made by [23]). Conversely, co-verbal behaviours of robots are generally added with mark-up languages such as SSML to extend the text input and trigger gestures. The SAIBA pipeline [5] enriched with PML [19] typically chains perception, decision and action processes that enrich associated representations of the system's state (i.e. PML, FML and BML) without considering fine-grained coordination constraints between input and output multimodal streams.

However few works (see below) have explored incremental dialog architectures that constantly monitor and potentially re-plan interactive multimodal behaviors.

3 COORDINATING SELF AND OTHERS' SPEECH AND GESTURES

Language and co-verbal gestures are widely accepted as an integral process of natural communication. Verbal communication is thus multimodal per se. This is true for verbal output: facial expression, iconic or emblematic gestures as well as gaze determine the addressee and may completely change the default meaning of an assertion, turning it to sarcasm, doubt or contempt. Multimodal patterns also include expectations: turn taking management elicits multimodal spatiotemporal patterns that bound behaviours of all parties [22].

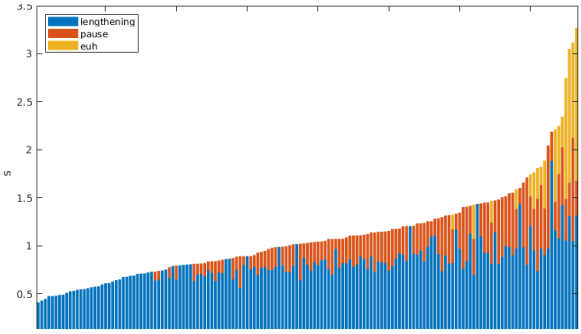

Figure 2: Speech samples awaiting for a co-verbal deictic 'That' gesture and desired attention from the partner to be performed in Put That There: we first observe final syllabic lengthening (blue), then pausing (red) and finally /um/ (yellow) . . . resulting in delays as long as 3s!

Multimodal recognition. Besides natural language, human communication often involves other modalities such as gaze, facial expression and hand gestures: deictic gestures often complement verbal information with physical location, relative placement .... Mapping multimodal utterances to semantic frames is still an open issue [3] but multimodal processing tames recognition errors.

Multimodal generation. Several open-loop systems trigger self-performed emblems, beat or deictic, metaphoric and iconic gesture. They use key timestamps linked with lexical items of synthetic speech, generated from text [7, 13] or pre-recorded speech [4]. Speech and gesture coordination is often performed asymmetrically by adjusting gesture kinematics to meet speech appointments [6]: such speech-driven coordination policy is rather difficult to set up with robots whose dynamics is less flexible than virtual avatars.

Coordinating speech generation with attention. Few close-loop systems have been attempting to incrementally coordinate speech production with listeners' visual focus of attention: the speech synthesis can be delayed at the end of phrases according to attentional demand. Such coordinative policy is complex: cognitive attention does not always manifest as eye contact, pausing should not be interpreted as floor releases ... [25]. Turn-taking management is such a cognitive ability that requires close-loop coordination of self and others' speech and gesture (both communicative and performative): Skanze et al [20] show that human-like coordination of speech, gaze and head movements in HRI facilitate real-time coordination and seamless turn-taking behaviors.

Predicting behaviors. In the near future, we expect dialog systems to be equipped with modules that compute expectations, predict others' behaviours and thus able to reconsider plans when strong deviations between expected and observed behaviours are experienced. Such modules have already been explored for diverse sensorimotor tasks such as driving [10], walking [8] or for predicting next activity [18]. Convolutional Neural Network front-ends are well-suited to detect spatiotemporal patterns in the flow of low-level perceptuomotor streams. Recurrent architectures (such as attention layers) may complement such specialized detectors/filters with context, including expectations.

4 WHICH CHALLENGES?

In order to promote the development of HRI architectures that enable **incremental processing** of speech and gestures of interlocutors while on-line planning and generating the robot's multimodal behaviours, **new HRI challenges should be designed,** that favor the cooperation between roboticists and specialists of multimodal interaction. Although simulated HRI may help the development of such incremental systems that constantly monitor sensorimotor loops and bootstrap learning of behavioural models, there is nothing like true task-oriented HRI situations with real users. The emblematic RoboCup (see https://www.robocup.org/) proposes such a meeting between technologists and end-users. It has four senior leagues: RoboCupSoccer, RoboCupRescue, RoboCup@Home and RoboCupIndustrial. Most tasks involve navigation (including following/heading persons) as well as interaction with objects of the environment (balls, valves, cans, knobs ...). Verbal interaction is often used to give or receive instructions and trigger action planning. But why not set-up a task where performative and communicative gestures are intricate and boost joint performance?

Sample challenges. The CRISSP team at GIPSA-Lab has mainly studied short-term task-oriented HRI (see figure 1) that aims at focusing on multimodal interaction while minimizing cognitive load: the roles of the agents are pre-determined (the robot is often instructing/coaching human partners), the task is rather simple. The challenge is to coordinate joint multimodal behaviours in order to have a smooth, seamless and efficient cooperation.

Evaluation and assessment. For task-oriented interaction scenarios, the time-to-completion remains the most objective evaluation criterion of efficiency. This is often complemented with posthoc questionnaires regarding the quality of users' experience, ratings of third parties as well as objective performance measures (duration of phases, number of turns...). These ratings are often biased by striking events that disrupt users' expectations, intentions and beliefs. We think that systems' assessment should benefit from online evaluation frameworks helping system developers in identifying misalignments and failures. Within this perspective, we proposed an online continuous assessment methodology [14] that mirrors evaluation currently performed by RoboCup judges: we ask third parties to review recorded interactions and to give a "yuck" response each time they experience an inappropriate behaviour from the robot. We showed that the "yuck" probability density function over time mirrors time-varying quality of conversational user experience.

Towards complex situations. The canvas proposed here deliberately covers up the Copernican revolution of automatic dialog management and state representation that such an incremental framework fosters. Task-based social interaction should be backed by planners [17] and dialog systems that provide context and reason on the task and the situation. The design of fully incremental dialog systems is of course a key issue for dealing with more complex open-domain HRI.

ACKNOWLEDGMENTS

We thank Sascha Fagel, Alaeddine Mihoub, Duc-Canh Nguyen, Frédéric Noël, Guillermo Gomez, Carole Plasson and Nathan Loudjani for their valuable contributions.

REFERENCES

[1] Laurence Devillers, Tatsuya Kawahara, Roger K. Moore, and Matthias Scheutz (Eds.). 2020. *Spoken Language Interaction with Virtual Agents and Robots (SLIVAR): Towards Effective and Ethical Interaction.*

[2] G Guillermo, P Carole, F Elisei, F Noel, and G Bailly. 2015. Qualitative assesment of a beaming environment for collaborative professional activities. In *European conference for Virtual Reality and Augmented Reality (EuroVR)*. 8 pages.

[3] Ting Han, Julian Hough, and David Schlangen. 2017. Natural Language Informs the Interpretation of Iconic Gestures: A Computational Approach. In *International Joint Conference on Natural Language Processing*. 134–139.

[4] Chien-Ming Huang and Bilge Mutlu. 2013. Modeling and Evaluating Narrative Gestures for Humanlike Robots.. In *Robotics: Science and Systems*. 57–64.

[5] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*. 205–217.

[6] Stefan Kopp and Ipke Wachsmuth. 2004. Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds* 15, 1 (2004), 39–52.

[7] Quoc Anh Le and Catherine Pelachaud. 2011. Generating co-speech gestures for the humanoid robot NAO through BML. In *International Gesture Workshop*. 228–237.

[8] Yuke Li. 2019. Which way are you going? imitative decision learning for path forecasting in dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*. 294–303.

[9] Hongyi Liu and Lihui Wang. 2018. Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics* 68 (2018), 355–367.

[10] Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. 2019. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7144–7153.

[11] Alaeddine Mihoub, Gérard Bailly, Christian Wolf, and Frédéric Elisei. 2015. Learning multimodal behavioral models for face-to-face social interaction. *Journal on Multimodal User Interfaces* 9, 3 (2015), 195–210.

[12] Roger K. Moore, Serge Thill, and Clémentine Vignal (Eds.). 2016. *Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR).*

[13] Victor Ng-Thow-Hing, Pengcheng Luo, and Sandra Okita. 2010. Synchronized gesture and speech production for humanoid robots. In *International Conference on Intelligent Robots and Systems*. 4617–4624.

[14] Duc Canh Nguyen, Gérard Bailly, and Frédéric Elisei. 2017. An evaluation framework to assess and correct the multimodal behavior of a humanoid robot in human-robot interaction. In *Gesture in Interaction (GESPIN)*. Posznan, Poland, 56–62.

[15] José Novoa, Jorge Wuth, Juan Pablo Escudero, Josué Fredes, Rodrigo Mahu, Richard M Stern, and Nestor Becerra Yoma. 2017. Robustness Over Time-Varying Channels in DNN-HMM ASR Based Human-Robot Interaction.. In *INTERSPEECH*. 839–843.

[16] Victor Paléologue, Jocelyn Martin, Amit Kumar Pandey, and Mohamed Chetouani. 2018. Semantic-based interaction for teaching robot behavior compositions using spoken language. In *International Conference on Social Robotics*. 421–430.

[17] Ronald PA Petrick and Mary Ellen Foster. 2013. Planning for social interaction in a robot bartender domain. In *Twenty-Third International Conference on Automated Planning and Scheduling*. Rome, Italy, 389–397.

[18] Tiziana Rotondo, Giovanni Maria Farinella, Valeria Tomaselli, and Sebastiano Battiato. 2019. Action Anticipation from Multimodal Data. In *International Conference on Computer Vision Theory and Applications*. Prague, Czech Republic, 154–161.

[19] Stefan Scherer, Stacy Marsella, Giota Stratou, Yuyu Xu, Fabrizio Morbini, Alesia Egan, Louis-Philippe Morency, et al. 2012. Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In *International Conference on Intelligent Virtual Agents*. 455–463.

[20] Gabriel Skantze. 2016. Real-time coordination in human-robot interaction using face and voice. *AI Magazine* 37, 4 (2016), 19–31.

[21] Dan Stowell, Angela Dassow, Ricard Marxer, Julian Hough, Roger K. Moore, Elisabetta Versace, Emmanouil Benetos, and Jessie Wand (Eds.). 2019. *Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR).*

[22] Kristinn R Thórisson. 2002. Natural turn-taking needs no manual: Computational theory and model, from perception to action. In *Multimodality in language and speech systems*. Springer, 173–207.

[23] Panagiota Tsarouchi, Sotiris Makris, and George Chryssolouris. 2016. Human–robot interaction review and challenges on task planning and programming. *International Journal of Computer Integrated Manufacturing* 29, 8 (2016), 916–931.

[24] Deng Yongda, Li Fang, and Xin Huang. 2018. Research on multimodal human-robot interaction based on speech and gesture. *Computers & Electrical Engineering* 72 (2018), 443–454.

[25] Zhou Yu, Dan Bohus, and Eric Horvitz. 2015. Incremental coordination: Attention-centric speech production in a physically situated conversational agent. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 402–406.