Multimodal analysis and recognition of social signals: application to social stance generation in virtual agents

T. Janssoone, 2015-2018
Supervisors: G. Richard (LTCI-CNRS), C. Clavel (LTCI-CNRS) and K. Bailly (ISIR)

We have worked on the first step of a methodology dedicated to deduce sequences of signals expressed by humans during an interaction. The aim is to link interpersonal stances with arrangements of social signals such as modulations of Action Units and prosody during a face-to-face exchange. The long-term goal is to infer association rules of signals. We plan to use them as an input to the animation of an Embodied Conversational Agent (ECA). We have illustrated the proposed methodology to the SEMAINE-DB corpus from which we automatically extracted Action Units (AUs), head positions, turn-taking and prosody information. We have designed the Social Multimodal Association Rules with Timing (SMART) algorithm. It proposes to learn the rules from the analysis of a multimodal corpus composed by audio-video recordings of human-human interactions. The methodology consists in applying a Sequence Mining algorithm using automatically extracted Social Signals such as prosody, head movements and facial muscles activation as an input. This allows us to infer Temporal Association Rules for the behaviour generation. We show that this method can automatically compute Temporal Association Rules coherent with prior results found in the literature especially in the psychology and sociology fields. The results of a perceptive evaluation confirms the ability of a Temporal Association Rules based agent to express a specific stance.

1. Context and objectives

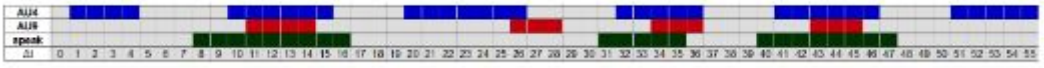
Embodied Conversational Agents (ECAs) can improve the quality of life in our modern digital society. For instance, they can help soldiers to recover from PTSD (Post Traumatic Stress Disorder) or help a patient to undergo treatment if they are empathic enough to provide support. The main challenge relies on the naturalness of the interaction between Humans and ECAs. With this aim, an ECA should be able to express different stances towards the user, as for instance dominance for a tutor or friendliness for a companion. This work proposes the SMART framework for the generation of believable behaviours conveying interpersonal stances.
Our work focuses on the scheduling of the multimodal signals expressed by a protagonist in an intra-synchrony study of his/her stance. Intra-synchrony refers here to the study of multimodal signals of one individual whereas the inter-synchrony studies the synchrony between two interlocutors. We focus on the sequencing that provides information about interpersonal stance as defined by Scherer as the "*characteristic of an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, coloring the interpersonal exchange in that situation (e.g. being polite, distant, cold warm, supportive, contemptuous)*". Indeed, the scheduling of non-verbal signals can lead to different interpretations: Keltner illustrates the importance of this multi-modality dynamics: a long smile shows amusement while a gaze down followed by a controlled smile displays embarrassment.

We are working on an automatic method based on a sequence-mining algorithm which aims to analyse the dynamics of the social signals such as facial expression, prosody, and turn-taking to deduce association rules with temporal information directly from social signals by transforming social signals into temporal events. The association rules are learnt from a corpus and will provide time-related information between the signal-based events in a sequence. However, one major difficulty to find these rules is that they are blended into each other due not only to the stance but also to other constraints such as identity, biomechanical constraints or the semantic contents of the given utterance. For instance, two persons can have a warm exchange but one frowns because

he/she is dazzled by the sun. Another example is that the AU 26, jaw drop, can signify surprise but can also be activated due to the speech production mechanisms.
After a survey of existing Temporal Constrained Systems solutions (Chronicle, Episode, etc), we focused on The Temporal Interval Tree Association Rule Learning (Titarl) algorithm because of its flexibility and its ability to express uncertainty and temporal inaccuracy of temporal events. Indeed, it can compute time relation as rules between events (before/after), negation and accurate time constraints such as

"If there is an event D at time t, then there is an event C at time t+5 ".

This temporal learning approach to find temporal associative rules from symbolic sequences allows to represent imprecise (non-deterministic) and inaccurate temporal information between social signals considered as events.
A temporal rule gives information about the relation between symbolic events with a temporal aspect. In our case, the events are the social signals (AUs, head nods, prosody, turn taking) considered as discrete events after a preprocessing step of symbolization.



For example, with the input of Fig.2, a temporal pattern could be:

If an event "activation of AU4" happens at time t while state Speak is active, then an event "activation of AU9" will be triggered between t+Dt and t+3Dt with a uniform distribution

which can be symbolized by the following rule

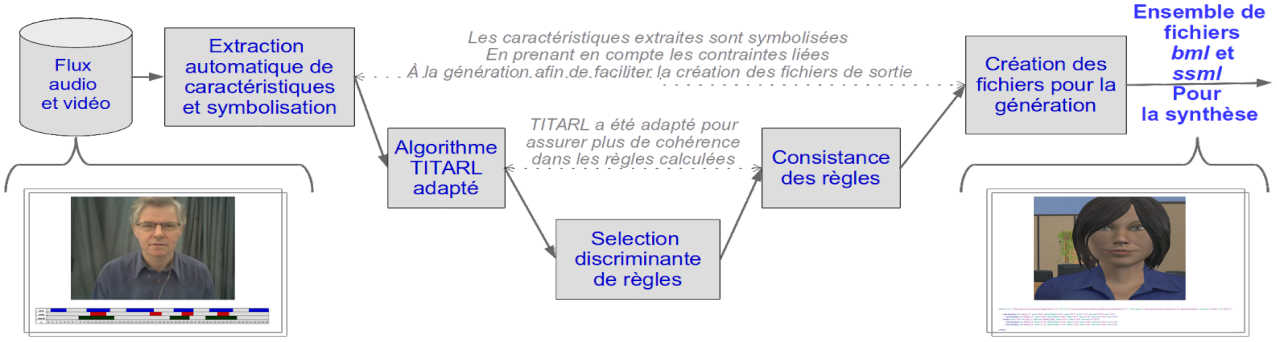
$$AU4_{\text{off to low}}^{\text{Speaking}} \xrightarrow{\Delta t, 3\Delta t} AU9_{\text{off to low}}$$

Some characteristics of a rule can be computed to validate its interest. If we look at the following rule r

$$A \xrightarrow{\Delta t_{min}, \Delta t_{max}} B$$

then the confidence of a rule is the probability of a prediction of the rule to be true. We are also interested in the support of a rule which is the percentage of events explained by the rule. Finally, TITARL ensures a good precision in the rule that is the temporal accuracy of the prediction, i.e., a low dispersal of the distribution of the events A (standard deviation) verifying the rule r.

We have modified TITARL and integrated it in a framework that aims to take audio-video files as an input and automatically generate files that are required to generate the behaviour of an Embodied Conversational Agent with a specific stance. The design of our framework is visible in the following figure



2. Main Results

We applied SMART on the SAL-SOLID SEMAINE database to illustrate our methodology. This corpus uses the Sensitive Artificial Listener (SAL) paradigm to generate emotionally coloured interactions between a user and a 'character' played by an operator. It proposes video and audio data streams of this Face-to-Face interaction where the operator answers with pre-defined utterances to the user's emotional state. We only focus here on the operator part where, for each session, he acts four defined roles, one by one, corresponding to the four quadrants of the Valence-Arousal space. Spike is aggressive, Poppy is cheerful, Obadiah is gloomy and Prudence is pragmatic. As a first step, for this study, we only focus here on two roles of the operator part, one friendly, Poppy, and one hostile, Spike. This represents 48 interactions of 34 minutes recording, 25 with Poppy, 23 with Spike, played by 4 different actors. This kind of data makes us restraint our study to the affiliation axis of the Argyle's theory of stance.

We performed studies to validate the extracted rules by comparing them to the results obtained in the literature in two steps. We retrieve that AUs corresponding to smile and cheek raiser (AU6, AU12) were linked to Poppy while brow lowerer (AU1/2 and AU4) were more linked to Spike.The literature explains that friendliness involves smile and cheek raiser while hostility is linked to brow lowerer.

Strengthened by the previous study, we conducted an evaluation of videos of an ECA generated from the best ranked Social Temporal Association Rules specific to a character. We processed the rules into BML files to use as an input of a virtual agent generation tool. The aim is to evaluate the perception of the agent's stance. We took the three best scored rules after 3 addition steps learned over the actor of the Semaine-SAL database in a listening status for each Poppy (friendly) and Spike (hostile). From these six rules, we got sequences of AU and head-nod evolutions with time information as we focus to the listener part. We also log the occurrences of each events verifying each rule to transpose them into BML files. These BML were used to generate video sequence with the virtual agent using the corresponding social signals. Hence we were able to synthesize an agent following these rules, with the timing of each transition set to the time of the highest occurrence. We used an agent to play each of this six rules and recorded its performances. We then used an online rating plateforme to evaluate our videos. Despite the very basic process of generation, rules characteristic of Poppy were perceived friendlier than the Spike's ones.

3. Communication

Associated Publication:

- **- Temporal Association Rules for Modelling Multimodal Social Signals,** *Authors* : Thomas Janssoone, *Publication date* : 015/11/9, *Conference* :Proceedings of the 2015 ACM on International Conference on Multimodal Interaction
- **- Des signaux sociaux aux attitudes : de l'utilisation des règles d'association temporelle,** *Authors* : Thomas Janssoone, Chloé Clavel, Kévin Bailly, Gaël Richard, *Publication date* : 2016/6/13, *Journal :* WACAI 2016, Workshop • Affect • Compagnon Artificiel • Interaction
- **- Using Temporal Association Rules for the Synthesis of Embodied Conversational Agents with a Specific Stance,** *Authors* : Thomas Janssoone, Chloé Clavel, Kévin Bailly, Gaël Richard, *Publication date :* 2016/9/20, *Conference* : International Conference on Intelligent Virtual Agents

Others :

- Participation at the GDR-ISIR : Poster (2015-12-02) and presentation (2016-11-15)
- ISSAS summer school, 2016. Price for the best student's project
- Scientific mediation at Cité des Sciences, thesis presentation and experiments with the public