Using temporal association rules for the synthesis of embodied conversational agents with a specific stance

Thomas Janssoone, Chloé Clavel, Kevin Bailly, Gael Richard

Using temporal association rules for the synthesis of embodied conversational agents with a specific stance

Thomas Janssoone, Chloé Clavel, Kévin Bailly, and Gaël Richard

{thomas.janssoone,kevin.bailly}@isir.upmc.fr
Sorbonne Universités, UPMC Univ Paris 06,CNRS UMR 7222, ISIR,F-75005, Paris, France
{chloe.clavel,gael.richard}@telecom-paristech.fr
Institut Mines-Télécom,Télécom-ParisTech CNRS-LTCI

Abstract. In the field of Embodied Conversational Agent (ECA) one of the main challenges is to generate socially believable agents. The long run objective of the present study is to infer rules for the multimodal generation of agents' socio-emotional behaviour. In this paper, we introduce 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 proposed 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.

Keywords: Multi-modal Social Signal, Sequence Mining, Signal Processing, Embodied Conversational Agent

1 Introduction

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 [1] 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 algorithm for the generation of believable behaviours conveying interpersonal stances.

To give ECAs the capacity to express emotions and interpersonal stances is one of the main challenges [2]. However, this field of research is thriving as more and more databases are available for the processing of Social Signals [3]. These databases are mainly audiovisual and provide monomodal or multimodal inputs to Machine Learning methods [4], [5], [6], [7]. Features such as prosodic descriptors or activations of facial

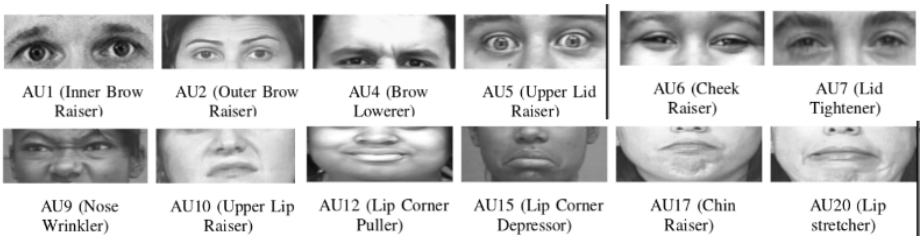

Fig. 1. Facial Action Unit locations, images are obtained from `http://www.cs.cmu.edu/~face/facs.htm`

muscles labelled as Action Units (AUs see Fig1) are extracted to recognize a social expression (emotion, stance, behavior ...). The data is usually labelled by an external observant who rates his/her perception of the ongoing interaction (*e.g* the levels of valence, of arousal, of antagonism, of tension ...). These annotations provide different classes for supervised machine learning algorithms.

This paper 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[8] 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[9] 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 present here 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. The focus is put on the processing of the input signals to find relevant sequences of temporal events through temporal association rules. To the best of our knowledge, this paper is the first attempt 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. From a long term perspective, the association rules will be dedicated to automatic temporal planning for the generation of ECA's stances that are believable. 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, bio-mechanical constraints or the semantic contents of the given utterance [10]. 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.

We detail how to process social signals such as AU and prosodic features as an input to a sequence-mining algorithm to find temporal association rules. The previous

approaches found in the literature will be detailed. Aferwards, the methodology will be explained: the considered multimodal social signals (Action Unit, head nod and turn taking), the process of the symbolization of the input signals into temporal events, the sequence mining algorithm and the scoring of the obtained Temporal Association Rules are justified. Then, the methodology is applied on a multimodal corpus in a human-agent interaction context to investigate the differences between someone cheerful and someone hostile during an interaction. These results will be reviewed and discussed by confronting them to the literature. A perceptive study of Temporal Association Rules based ECA is also detailed and discussed. Finally, perspectives and future leads will be presented.

2 Related work

The links between social signals and interpersonal stances have been studied during the last decades [11] with goals such as the detection of the user's/human's stance or the generation of believable interpersonal stances for ECAs. To do so, humans' expressions of these stances were studied.

First, qualitative studies were made such as Allwood et al. [12] where verbal and gestural feedbacks during dialogues between a travel agent and customers were investigated. The relationship between prosody and gestures was underlined in this particular context.

In qualitative studies about first impression, Cafaro et al. [13] show how the observant's feeling of the stance of a virtual character is impacted by nonverbal immediacy cues. They underline that proximity has no effects on judgements of friendliness. In [14], a cluster approach shows the link between head-nod, head-shake and affect labels made either with audiovisual files or visual only. They show a strong affective meaning of the nod and the shake and underline the limit of the inter-rater agreement due to the verbal context available for one party. These approaches use statistical tools to link social signals and perceptions of stances.

Lately, machine-learning algorithms were used like in the study of Lee and Marsella[15] about how head-nod magnitude and eyebrow movements evolve while speaking. Participants were asked to rate their immediate feeling of a virtual agent speaking while making head nods and eyebrow movements. Three learning algorithms (Hidden Markov Model, Conditional Random Fields and Latent-Dynamic Conditional Random Fields) were compared to model head nods and eyebrow movements. However, even if they improve the recognition, this model did not improve the generation of realistic stances, maybe due to the hypothesis tested. Ravenet et al. [16] create a corpus of ECAs postures according to several stances. They develop a Bayesian model to automatically generate stance which however does not take into account the temporal aspects of the signals used to express a stance.

Finally, sequence-mining algorithms have been explored to find input for machine-learning based generation of agent. For instance, Martinez et al.[17] and then Chollet et al. [18] explain how to use them to find simple sequences of non-verbal signals associated to social stances. Martinez et al. [17] use the particular context of video gaming to link these feature samples to emotions such as frustration. They use the Generalised Sequence Pattern (GSP) algorithm on physiological signals to predict the player's affective state. Yet, the obtained sequences are not used for gen-

eration. Chollet et al. [18] also used GSP algorithm to extract sequences of manually annotated non-verbal signals characterizing different interpersonal stances. Hence, the GSP algorithm extracts sequences of events without temporal information *i.e.* it can only find that one event happens after another. Then, a model for the expression of a particular stance by an ECA was built to select the most appropriate sequence. Although these studies have proposed an analysis of social signal sequences, they do not consider temporal information. However, such information may change the interpretation of social signals sequences e.g. a long smile versus a short one as shown in [9].

3 Our approach

In the same vein, the chosen approach takes advantage of a sequence mining algorithm. The new contributions of our approach rely, firstly, on adding temporal information and, secondly, on directly processing the audio-visual input signals. To do so, the signals are transformed to be seen as temporal events (Section 3.1) that are the inputs of the sequence mining algorithm and we choose to use the Temporal Interval Tree Association Rule Learning algorithm (TITARL) as sequence mining algorithm (Section 3.2). We adapt the TITARL algorithm and embedded its new version into our framework Social Multimodal Association Rules with Timing (SMART) that we detail below (Section 3.3).

3.1 Feature extraction

We choose to focus on a set of social signals composed by *facial AUs activation* (see Fig1), *the head pose* and information such as *dialogic events*. We detail here the process that is used to compute these descriptors.

The 3D head pose was estimated with Intraface [19], a fully automatic face tracker. Its outputs are the pitch, the yaw and the roll of the head of the actor present in the video. We use these values as descriptor after a moving average smoothing over a 3 frames window and we cluster them in 10 degrees group. We then create events when the head passes from one 10 degrees group to one of the next 10 degrees group. Hence we keep the continuity of the original signal in our new symbolic temporal events.

Dialogic events correspond to events related to turn-taking activity. They indicate whether the human is listening or speaking so we get the action of *start speaking* and *end speaking* as new events. This information is supplied by the manual transcript provided with the studied corpus (see Section 4). Dialogic information is also used to annotate the state (speaking or listening mode) of other events such as AUs or head nod.

Prosodic features were extracted using Prosogram, a program developed by Mertens [20] which aims to provide a representation of intonation as perceived by a human listener. We choose Prosogram among other tools for automatic prosodic annotation because of its phonetic approach that better reflects the human perception than the other approaches. Indeed, all pitch movements cannot be perceived by the human ear and, as our long term goal is stance generation, we focus on signals that will play a part in perception.
Furthermore, Prosogram proposes an automatic segmentation of the audio files into

syllabic like nuclei and computes global prosodic parameters such as speaker pitch range. Then, it transforms these into an approximation of perceived pitch patterns. This bottom-up approach does not need additional information such as annotation or training, and then avoids the risk of bias.
For this study, we compute for each nuclei the mean f_0, and its variation, the peak of intensity, and the shape of the pitch (rises, falls, ...). Then, we merge the nuclei information of the shape at the word level to obtain the shape of the pitch inside the word. The timing information of each word is provided by the transcript. The three other features, f_0, its variation and the peak of intensity, as they are continuous, are turned into symbolized temporal events with the SAX symbolization process [21].

The Action Units were automatically detected using the solution proposed by Nicolle and al.[22]. An exponential smoothing was applied with $\alpha = 0.7$ on this continuous output to reduce the noise of the detection. Then, the AUs were symbolized as three folders: inactivate, low activation and high activation. For the study presented in 4, we focus on AUs corresponding to smile and eyebrow movements (see Fig 1). AU 1 and 2 are grouped and describe brow raising. AU 4 describes brow lowering. AU 6 describes cheek raising. AU 12 describes lip corner pulling. When two AUs are grouped, the value kept is the maximum value of each.
We consider as events the variation of AU activation like for example AU6 disabled to low activation will be an $AU6_{\text{off to low}}$ event or AU12 from low to high activation an $AU12_{\text{low to high}}$ event. We also provide for each event the state of the person, listening or speaking, thanks to the dialogic event formely detailed.

3.2 Sequence Mining algorithm

After a survey of existing Temporal Constrained Systems solutions (Chronicle, Episode, etc), we focused on The Temporal Interval Tree Association Rule Learning (Titarl) algorithm [23] 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*

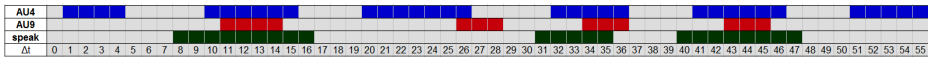

Fig. 2. example of social signal input for TITARL

"activation of AU4" happens at time t while state Speak is active, then an event "activation of AU9" will be triggered between t+Δt and t+3Δt with a uniform distribution which can be symbolized by the rule following in (1):

$$\underset{Speaking}{AU4_{\text{off to low}}} \xrightarrow{\Delta t, 3\Delta t} AU9_{\text{off to low}} \tag{1}$$

Δt represents here a time-step due to the training data such as the video frame rate. A rule is composed by a head, here the $AU9_{off\ to\ low}$, a tree of temporal constraints, here the

$$\underset{Speaking}{AU4_{\text{off to low}}},$$

a temporal distribution, here the $\Delta t, 3\Delta t$.

Some characteristics of a rule can be computed to validate its interest. If we look at the following rule r defined in eq.(2):

$$A \xrightarrow{\Delta t_{min}, \Delta t_{max}} B \tag{2}$$

then the confidence of a rule is the probability of a prediction of the rule to be true (see 3a). We are also interested in the support of a rule which is the percentage of events explained by the rule (see 3b). 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.

$$confidence = P(B(t')|A(t)), t' - t \in [\Delta t_{min}, \Delta t_{max}] \tag{3a}$$

$$support = \frac{\#\, B, \exists A\ such\ that\ (A \rightarrow B)\ true}{\#\, B} \tag{3b}$$

$$precision = \frac{1}{\text{std}([\text{t'-t, } \exists \text{ A, B, } (A(t) \rightarrow B(t')) \text{ true}])} \tag{3c}$$

For example, for the rule in 1, the confidence will be 1.0 and the support 0.75.
More details about the TITARL algorithm can be found in [23].

3.3 Adjustment of the TITARL algorithm into SMART

Introduction to the TITARL algorithm The structure of the TITARL pipeline, shown in the Fig.3, has two major parts. With the set of input events, characterized by a name, a session, a character and a time, a first process will generate simple rules with a structure like in 2 and a very large temporal distribution. These rules have a high confidence and support but a very low precision. The three steps of TITARL are recursively applied: division of the rule, refinement and addition of conditions.
The division step will produce more accurate rules by dealing with co-occurence of events. For example, if we have an event A at time t, an event B at time $t + 5\Delta t$ and an event B at time $t + 15\Delta t$, we can have the rule $A \xrightarrow{\mathscr{U}_{5\Delta t, 15\Delta t}} B$ or the two rules $A \xrightarrow{\mathscr{U}_{5\Delta t}} B$ and $A \xrightarrow{\mathscr{U}_{15\Delta t}} B$. This division step will choose between the two possibilities.

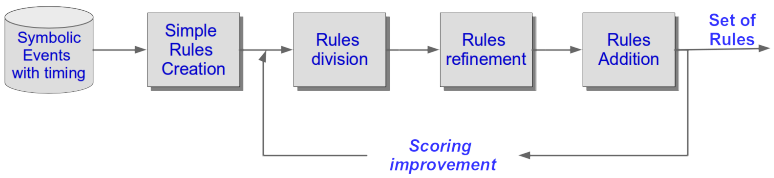

Fig. 3. pipeline of the TITARL algorithm

The refinement of a rule aims at increasing its precision: it observes the temporal distribution of the events verifying the rule and decreases their variance with a threshold on its histogram of distribution.
The addition of a condition to a rule is done to maximize the information gain. To do so, given a rule $A \rightarrow B$, the algorithm observes all the simple rules like $B \rightarrow C$. It will combine them as $A \rightarrow B \rightarrow C$ if its number of occurrences remains over a fixed threshold. This step will stop in case of a loop in the chain of events.
These last steps are carried on while the product of the confidence, the support and the precision remains over a threshold.

The original TITARL algorithm had some limitations and we detail here how we handled them into the SMART framework that can be seen in fig4.

Discriminative selection of the Association Rules: A score shown in eq.4a was defined as a combination of the confidence, the support and the size of the temporal interval of the rules to rank the computed Temporal Association Rules. This score was introduced by Guillame-Bert in [23] and tried out on artificial dataset and a 'Home

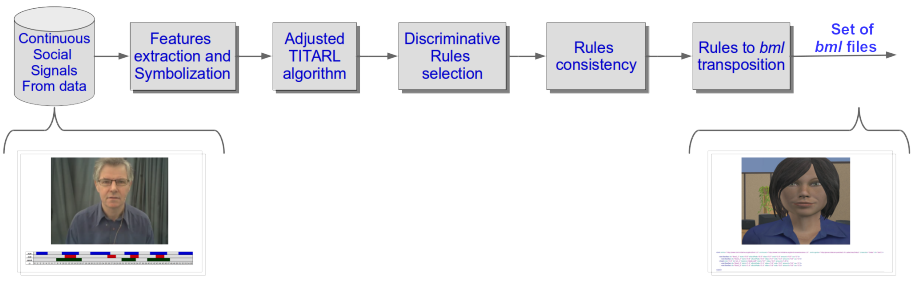

Fig. 4. pipeline of the SMART algorithm

Activities' dataset.

$$score = \frac{conf_r^4 . supp_r^2}{t_{max} - t_{min}} \tag{4a}$$

$$freq_{\text{character}_C}(r) = \frac{\text{number of occurence of rule r for character C}}{\text{total length of the data}} \tag{4b}$$

$$freqRatio(r, \text{character}_C, \text{character}_D) = \frac{freq_{\text{character}_C}(r)}{freq_{\text{character}_D}(r)} \tag{4c}$$

This score reflect the relevance of a rule in a general context. However, for our purpose, a rule can have a high score without being linked to a stance. As we also want to know the accuracy of the rule with a specific stance, we also defined the frequency of a rule. As described in eq.4b, it consist of the ratio between the number of occurrences of the rule and the length of the session. Indeed, as it will be described below, the learning is done on video of actor playing characters with a very specific stance. To differ rule specific to a stance played by a character, we can then use the ratio of frequencies detailed in eq.4c. This ratio of frequencies enables to prune the rules linked to a specific stance form others. For instance, if the ratio is high between a friendly video and a hostile one, the rule may be relevant to generate a friendly stance. In the mean time, rules corresponding to jaw movements due to the speech production mechanism are detected with a close to one frequency ratio. We can then not consider them as a stance relevant rule. This part is made in the *Discriminative rule selection* of the figure 4.

Rule consistency for the generation of relevant stance for ECAs: An other issue with TITARL was to handle the consistency of the rule. The association were originally made between all the signals. For one social signal, this could lead to irrelevant rules which provide no information when the transition are lost. For example, we could have rules like in the rule in eq.5

$$Event_{\text{off to low}} \xrightarrow{\Delta t_{min};\Delta t_{max}} Event_{\text{off to low}} \tag{5}$$

but we do not know how and when the Event went from low to off value again. With our generation perspective, this information is essential. To deal with this, we modify TITARL, the simple rule part to be specific, and we design two computational strategies: intra-signal and inter signal. This corresponds to the *Adapted TITARL* box in the fig 4.For instance, intra signal applies to two consecutive changes for the same AU while inter signal will be an AU and a head movement or two different AUs.
For intra signal, we compute the simple rules only with the previous occurrence of the same signal. Hence, the consistency is assured. For inter signal, we keep all the previous occurrence, so the original TITARL design. This reduces inaccurate rules computation by improving the consistency of all the transitions in the Temporal Association Rule. Hence, the rule in eq.5 cannot be consider and is replaced by the following one:

$$Event_{\text{off to low}} \xrightarrow{\Delta t_{min_1};\Delta t_{max_1}} Event_{\text{low to off}} \xrightarrow{\Delta t_{min_2};\Delta t_{max_2}} Event_{\text{off to low}} \tag{6}$$

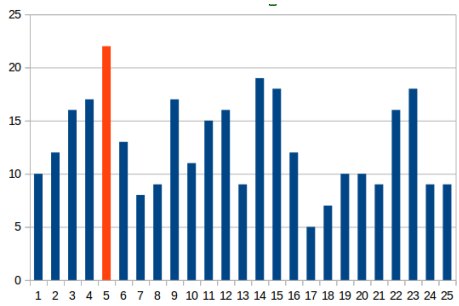

Fig. 5. example of distribution of occurrences for a rule

Behavior Markup Langage file generation for ECAs: The last contribution we made in the SMART framework is the transposition of the Temporal Association Rule into Bahavior Markup Langage files (*BML*). The Behavior Markup Language, or BML, is an XML description language for controlling the verbal and nonverbal behavior of (humanoid) embodied conversational agents (ECAs). A BML block describes the physical realization of behaviors (such as speech and gesture) and the synchronization constraints between these behaviors. This is the last part of the framework in fig 4.
Through this BML, we can provide the timing used to control the social signals expressed by the ECA during an animation. To do so, we also log the occurrences of each events verifying each rule. In this first version of the SMART framework, we simply use the timing with the most occurences for the transition (red bar at $\Delta_t = 5$ in fig.5. In a future release, we could also use the whole distribution of events verifying the rule to diversify the transition and so have more diverse synthesis of a stance.

4 Social Signals as Temporal Events: application for the study of the SEMAINE-db corpus

4.1 The SEMAINE-db corpus

We applied TITARL on the SAL-SOLID SEMAINE database [24] 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 $3-4$ 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 theroy of stance

[25]. The characters of Poppy and Spike overact it very well and this allows us a first validation of our model.

4.2 First study: comparison to the literature

We performed a first study to validate the extracted rules by comparing them to the results obtained in the literature in two steps. The first one focuses on the rules combining specific AUs, the second one focuses on the rules combining the AUs with the prosodic events.

As a first step, we choose to consider here more specifically AUs corresponding to smile and cheek raiser (AU6, AU12) and brow lowerer (AU1/2 and AU4) and to test TITARL on these specific social signals. Indeed, we want here to compare the connections highlighted in [26], [16] on an ECA study with our results. These papers explain that friendliness involves smile and cheek raiser while hostility is linked to brow lowerer.

In Table1, association rules are shown with their confidence, support, score and frequency ratio. We choose to show the rules with the highest score with a discriminant frequency ratio. We can see that Poppy, who acts as friendly, is more likely to smile than Spike. Actually, the low frequency ratio leads to think that this is due to the speech production mechanism. This hypothesis is strengthened by the fact that a large part of AU6 and AU12 activation for Spike are while speaking. For the brows, Spike frown more, especially while speaking but a noteworthy result is about Poppy frowning while listening. This can be considered as a backchannel to notice the speaker of Poppy's interest in the conversation.
These results are not only consistent with the literature but also able to provide temporal information and confidence. Indeed, the empirical and theoretical research have shown

	rule ($body \xrightarrow{\Delta t_{min};\Delta t_{max}} head$)	**confidence**	**support**	**score**	**frequency ratio**
Poppy	$AU6_{\text{off to low / listening}} \xrightarrow{0.0s;0.2s} AU6_{\text{low to off/ listening}}$	0.64	0.63	3.10^{-2}	2.09
Poppy	$AU12_{\text{off to low / listening}} \xrightarrow{0.0s;0.2s} AU12_{\text{low to off/ listening}}$	0.50	0.51	8.10^{-3}	3.78
Spike	$AU4_{\text{low to high / speaking}} \xrightarrow{0.0s;0.2s} AU4_{\text{high to low / speaking}}$	0.76	0.81	1.10^{-1}	1.62
Poppy	$AU4_{\text{off to low / listening}} \xrightarrow{0.0s;0.2s} AU4_{\text{low to off/ listening}}$	0.71	0.71	6.10^{-2}	2.07
Spike	nuclei $f0_{\text{large decrease}} \xrightarrow{0;0.9s} AU1+2_{\text{off to low}} \xrightarrow{0;0.3s} AU1+2_{\text{low to off}}$	0.82	0.17	3.10^{-4}	0.88
Poppy	word shape of $f0_{down} \xrightarrow{0;0.9s} AU4_{\text{low to off}}$	0.53	0.57	1.10^{-5}	1.44
Spike	word shape of $f0_{\text{up and down}} \xrightarrow{0.1;0.8s} AU4_{\text{off to low}} \xrightarrow{-0.1;0.3s} AU4_{\text{low to off}}$	0.74	0.01	2.10^{-5}	0.88
Poppy	start speaking $\xrightarrow{0;0.6s} AU1+2_{\text{low to high}} \xrightarrow{0;0.3s} AU1+2_{\text{high to low}}$	0.74	0.57	1.10^{-3}	2.43

Table 1. Sample of results with their confidence, support and score. The two first rows are linked to the smile and the two following are linked to the eyebrow. The four last ones refers to the second study and link the prosody to the brow movements.

that friendly stances imply frequent smiles while frowning are perceived as threatening and hostile. Our study enables us to identify more precisely the duration of the social signals. Such information is essential to synthesize the stance of an ECA.

In a second phase, we aim here to validate the process of this methodology combining audio and visual information. To do so, a priori rules were deduced from the literature, confronted to the ones obtained here and then discussed. Guaïtella and al. [27] present an experimental investigation on the link between eyebrow movements to voice variations and turn-taking. It shows that peak of the contour of the fundamental frequency are associated with the observed eyebrow movements but it appears that the reverse is false. It also measure the link between eyebrow-movement and start-or-end-of-vocalization as "eyebrow movements act as marks of a new speaking turn". Roon et al. [28] complete these results by underlining that the relation between eyebrow and head position is much closer than the one between eyebrow and speech.

We investigate here these relations by computing with SMART the association rules for eyebrow ($AU1+2$ and $AU4$), prosody (f_0 and variation) and turn taking. Some computed rules are shown in Table 1 for the shape of the f_0 at the word level and the brow activation considering Poppy and Spike. We can see that the prosodic shape of a word is followed by a similar activation of the brows. It may be explained by the expressiveness of these two characters and the use of brows to emphasise the Irish/English dynamic accent. However, no significant difference was found between Poppy and Spike, especially for the turn taking. This may be explained by the metric that is not appropriate for such a specific event. We plan to improve it with information retrieval techniques to find relevant keyword in documents. Finally, a remarkable result appears for the turn-taking of Poppy. An intense raise of the brow often follows the start of speaking that may be used to improve the bond while taking the floor and, so, improve the friendliness.

4.3 Second study: perspective evaluation with a focus on AUs and headnods

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 [29]. The aim is to evaluate the perception of the agent's stance. As we were in a generation process, we restrained the set of input signals to head nod movements and AUs we can control on the ECA as explained in section 3.1. For instance, AU_9 corresponding to "nose wrinkler" was not implemented so we did not use it.

The design of the study was the following: 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 signales. Hence we were able to synthesized 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.

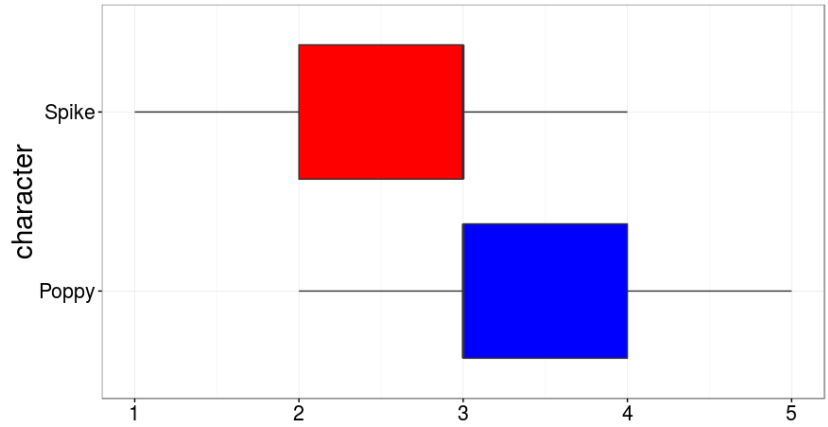

Fig. 6. Boxplot of the evaluation of the affiliation

We then used an on-line platform to get 60 ratings of each of the 6 videos. 97 judgements were done by 62 participants and each participant was asked to rate his/her feeling of the affiliation (hostile/friendly) of the ECA and his/her confidence in his/her judgement with two five-point Likert-like scales. For instance, the affiliation rating went from 1 corresponding to *very hostile* to 5 *very friendly* through 2 *hostile*, 3 *neutral* and 4 *friendly*.

We first analyse the overall results of the ratings of the poppy-based (friendly) video and the spike-based (hostile) video. The summary of the answer can be seen in the table 2. We can see that the mean of the answer for Spike is 2.5 and the third quartile is 3 that means that 75% of the evaluation rate between 1 and 3 which was the expected results. Likewise, we can see for the one about Poppy that the mean is 3.367 and 75% of the answers were between 3 and 5. For the rest of this study, we followed the statistical advise from [30].

As Shapiro-Wilk test indicates that the answers did not followed a normal distribution (both p-value$< 10^{-16}$), we ran a Mann-Withney's U test to evaluate the difference in the response and we found a significant effect of Group (p$= 9.10^{-5}$). This confirms that the results of the Poppy based evaluation are higher than the Spike one. The results are shown as a boxplot in Fig6. This confirm that a video based on a Temporal Association

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
Poppy	2.000	3.000	3.000	3.367	4.000	5.000
Spike	1.000	2.000	3.000	2.5	3.000	5.000

Table 2. Summary of the global answer over Poppy and Spike based video.

Rule caracteristic of Poppy is seen friendlier than a Spike one, despite this very basic synthesis process.

5 Conclusion and future work

This paper presents a methodology to compute temporal association rules from automatically extracted multimodal signals. The focus is put on the processing of these signals to use them as an input of this sequence mining algorithm. The studies show that such methodology can be used to identify interesting sequences of social signals. The adaptation of the score and the use of frequency of occurences give is an important new feature. This combination manages to discriminate rules du to bio-mechanical constraints (i.e. speech production) to others relevant to a stance. However, we believe that this pruning can be still improved in future works.
The studies also shows the ability of SMART to find temporal information associated to specific stances. It is efficient to retrieve relevant rules detailed in the literature with temporal information, confidence and support. The two studies validated this approach but also open the perspective for future developments. The prosodic aspect remains challenging but we plan to use SMART to retrieve contours of sentences. We are now struggling with subtle issues such as dynamic accent as presented in the studies.

6 Acknowledgement

This work was performed within the Labex SMART supported by French state funds managed by the ANR within the Investissements d´Avenir programme under reference ANR-11-IDEX-0004-0.

References

1. K. Truong, D. Heylen, M. Chetouani, B. Mutlu, and A. A. Salah. Workshop on emotion representations and modelling for companion systems. In *ERM4CT@ICMI*, 2015.
2. A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 2009.
3. A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *Affective Computing*, 2012.
4. O. Rudovic, M. A Nicolaou, and V. Pavlovic. 1 machine learning methods for social signal processing.
5. A. Pentland. Social dynamics: Signals and behavior. In *ICDL*, 2004.
6. G. Sandbach, S. Zafeiriou, and M. Pantic. Markov random field structures for facial action unit intensity estimation. In *ICCVW*, 2013.
7. A. Savran, H. Cao, A. Nenkova, and R. Verma. Temporal bayesian fusion for affect sensing: Combining video, audio, and lexical modalities. 2014.
8. K. R. Scherer. What are emotions? and how can they be measured? *Social science information*, 2005.
9. D. Keltner. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 1995.

10. E. Bevacqua and C. Pelachaud. Expressive audio-visual speech. *Computer Animation and Virtual Worlds*, 2004.
11. Q. Fu, R. op den Akker, and M. Bruijnes. A literature review of typical behavior of different interpersonal attitude. *Capita Selecta HMI, University of Twente*, 2014.
12. J. Allwood and L. Cerrato. A study of gestural feedback expressions. In *First nordic symposium on multimodal communication*, 2003.
13. A. Cafaro, H. H. Vilhjálmsson, T. Bickmore, D. Heylen, K. R. Jóhannsdóttir, and G. S. Valgardsson. First impressions: Users judgments of virtual agents personality and interpersonal attitude in first encounters. In *IVA*, 2012.
14. R Cowie, H Gunes, G McKeown, J Armstrong, and E Douglas-Cowie. The emotional and communicative significance of head nods and shakes in a naturalistic database.
15. J. Lee and S. Marsella. Modeling speaker behavior: A comparison of two approaches. In *IVA*, 2012.
16. B. Ravenet, M. Ochs, and C. Pelachaud. From a user-created corpus of virtual agents non-verbal behavior to a computational model of interpersonal attitudes. In *IVA*, 2013.
17. H. P. Martínez and G. N. Yannakakis. Mining multimodal sequential patterns: a case study on affect detection. In *ICMI*, 2011.
18. M. Chollet, M. Ochs, and C. Pelachaud. From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In *IVA*, 2014.
19. Xiong X and De la Torre F. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.
20. P. Mertens. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Speech Prosody 2004, International Conference*, 2004.
21. J. Lin, E. Keogh, L. Wei, and S. Lonardi. Experiencing sax: A novel symbolic representation of time series. *Data Min. Knowl. Discov.*, 2007.
22. J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *ICMI*, 2012.
23. M. Guillame-Bert and J. L. Crowley. Learning temporal association rules on symbolic time sequences. In *ACML*, 2012.
24. G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing*, 2012.
25. Michael Argyle. *Bodily communication*. Routledge, 2013.
26. M. Ochs and C. Pelachaud. Model of the perception of smiling virtual character. In *AAMAS*, 2012.
27. I. Guaïtella, S. Santi, B. Lagrue, and C. Cavé. Are eyebrow movements linked to voice variations and turn-taking in dialogue? an experimental investigation. *Language and speech*, 2009.
28. K. D. Roon, M. K. Tiede, K. M. Dawson, and D.H. Whalen. Coordination of eyebrow movement with speech acoustics and head movement. *ICPhS*, 2015.
29. F. Pecune, A. Cafaro, M. Chollet, P. Philippe, and C. Pelachaud. Suggestions for extending saiba with the vib platform. *Proceedings of the Workshop on Architectures and Standards for IVA*, 2014.
30. H. Motulsky. *Intuitive biostatistics: a nonmathematical guide to statistical thinking*. Oxford University Press, USA, 2013.