Des signaux sociaux aux attitudes : de l'utilisation des règles d'association temporelle

Thomas Janssoone
Pierre et Marie Curie University - ISIR
Paris, France
thomas.janssoone@isir.upmc.fr

Chloé Clavel
Institut Mines-Télécom,Télécom-ParisTech
CNRS-LTCI
Paris, France
chloe.clavel@telecom-paristech.fr

Kévin Bailly
Pierre et Marie Curie University - ISIR
Paris, France
kevin.bailly@upmc.fr

Gaël Richard
Institut Mines-Télécom,Télécom-ParisTech
CNRS-LTCI
Paris, France
gael.richard@telecom-paristech.fr

ABSTRACT

Dans le domaine des Agents Conversationnels Animés (ACA), l'un des principaux défis est de générer des agents réalistes. Le but à long terme de l'étude présentée dans cet article est de générer des règles pour la synthèse d'attitudes d'ACA. Nous proposons ici l'analyse de corpus audio-vidéo d'interactions humain-humain afin de déterminer ces règles. La méthodologie proposée utilise pour cela un algorithme de sequence mining sur des signaux sociaux extraits automatiquement comme les expressions faciales ou la prosodie. Nous montrons ici que cette méthode permet de calculer des Règles d'Associations Temporelles de manière autonome et que celles ci sont cohérentes avec des résultats à priori obtenus dans la littérature en psychologie et sociologie.

1. INTRODUCTION

L'utilisation d'Agents Conversationnels Animés (ACA) est une solution envisagée afin d'améliorer la qualité de vie de nos sociétés modernes. Par exemple, un psychiatre virtuel peut aider des soldats à récupérer d'un trouble de stress post-traumatique ou un autre agent virtuel peut encourager des malades à bien suivre leurs traitements tout en leur permettant de rester autonomes [1]. Le principal défi est de proposer une interaction naturelle entre des humains et des ACAs. Pour cela, ce dernier doit être capable d'exprimer différentes attitudes envers un utilisateur, comme de la dominance pour un tuteur ou de la bienveillance pour un compagnon.

L'un des principaux défis du domaine des agents virtuels est de donner aux ACAs cette capacité d'exprimer des émotions et des attitudes sociales [2]. Néanmoins, ce champ de recherche est en plein développement et de nombreux corpus sont disponibles pour l'étudier [3]. Ces corpus sont des bases de données principalement composées de flux audio et vidéo, fournissant ainsi des données d'entrée, monomodales ou multimodales, pour des méthodes de machine-learning [4]. Des caractéristiques comme des descripteurs prosodiques ou les activations des muscles faciaux décrites comme des Action Units (AUs voir image 1) en sont extraites afin de reconnaître des expressions sociales (émotions, attitudes, comportements, ...). Ces données sont généralement annotées par un (ou plusieurs) observateur extérieur qui va indiquer sa perception de l'interaction (*e.g.* degré d'antagonisme, de tension ou d'excitation ...). Ces annotations fournissent différentes classes pour les utiliser en entrées d'algorithmes de machine-learning.

Cet article se concentre sur la partie attitudes sociales des interactions, attitudes sociales au sens de Scherer [5] comme la "caractéristique d'un style affectif qui se développe spon-

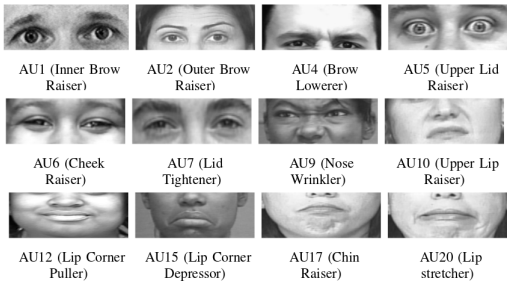

Figure 1: position des Facial Action Unit, images obtenues sur http://www.cs.cmu.edu/~face/facs.htm

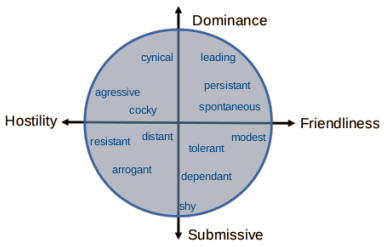

Figure 2: Circomplexe interpersonnel des dimensions de l'attitude. Les adjectifs illustrant des positions proviennent de [6]

tanément ou est stratégiquement employé lors d'une interaction avec une personne ou un groupe de personnes, colorant l'échange interpersonnel dans ce contexte (e.g. être poli, distant, froid, chaleureux, compassionnel, dédaigneux)". Argyle en propose une représentation selon deux axes, la dominance et l'appréciation, qui permettent de définir un circomplexe où une attitude correspond à une position dans cet espace bidimensionnel dont une illustration est présentée sur la figure 2.

Nous nous sommes en particulier intéressés à la dynamique de ces signaux car elle va apporter des informations nécessaires à la compréhension d'une attitude. Keltner [7] en illustre l'importance avec l'exemple suivant : un long sourire va être révélateur d'amusement tandis qu'un regard fuyant suivi d'un sourire contraint sera signe d'embarras.

Pour modéliser cet aspect dynamique des attitudes, une méthode, totalement automatique est ici proposée, basée sur un algorithme de sequence-mining afin d'analyser l'enchaînement des signaux sociaux tels que les expressions faciales, les mouvements de tête, la prosodie ou la prise de parole. L'idée est donc de considérer les modulations de ces signaux comme des séquences d'événements temporels symboliques afin d'en déduire des règles d'associations. Ces règles d'associations vont contenir l'information temporelle entre les occurrences de ces signaux mais aussi des données sur la distribution de ces séquences dans une interaction. Cela permettra, sur le long terme, de générer un modèle d'attitudes pour un ECA basé sur le planning de ces règles.

Cependant, une des principales difficultés pour trouver ces règles d'association est qu'elles sont entremêlées avec d'autres règles liées à l'identité, à des contraintes biomécaniques ou au contenu sémantique du discours tenu [8]. Par exemple, deux personnes peuvent avoir un échange très chaleureux mais l'une d'elle, éblouie par le soleil, fronce les sourcils. Dans un autre cas, l'AU 26 correspondant à la baisse de la mâchoire peut signifier la surprise mais peut aussi être activée lors du mécanisme de production de la parole.

La suite de cet article présente ce traitement des signaux sociaux afin de les utiliser en entrée d'un algorithme de sequence-mining. Dans un premier temps, les approches précédentes trouvées dans la littérature sont détaillées, puis la méthodologie est présentée. Les signaux sociaux multimodaux considérés (sourire, sourcils, descripteurs prosodiques, tour de parole) sont introduits avec un focus mis sur leurs transformations en événements temporels symboliques. L'algorithme de sequence-mining utilisé et le score des règles obtenues sont ensuite justifiés. Enfin, deux études illustrent l'utilisation de cette méthode sur un corpus multimodal avant une discussion sur les perspectives et les développements futurs.

2. ÉTAT DE L'ART

Cette relation entre les signaux sociaux et les attitudes sociales a été étudiée durant les dernières décennies [9] avec deux principaux buts : soit pour détecter l'attitude d'un utilisateur, soit pour générer des attitudes crédibles pour animer un ECA. La méthode courante est d'observer des humains exprimant ou réagissant à différentes attitudes.

Dans des études qualitatives, Cafaro et al.[10] étudient la première impression qu'a un observateur de l'attitude d'un personnage virtuel et comment celle-ci est modifiée selon différents signaux non-verbaux. Ils insistent en particulier sur le fait que la proximité entre l'observateur et l'agent n'a pas d'impact sur le jugement de gentillesse.

Cowie et al.[11] proposent une approche par clusters pour montrer le lien entre les mouvements de tête (selon l'axe de rotation) et des labels sur l'affect définis avec la vidéo seule ou avec la vidéo et le son. Ils montrent une forte corrélation entre l'affect (positif ou négatif) et le sens du mouvement. Ils soulignent également la limite entre la cohésion des annotations et le contexte verbal fourni seulement à une partie des annotateurs. Ces approches utilisent l'outil statistique pour faire le lien entre signal social et perception de l'attitude.

Par ailleurs, des algorithmes de machine-learning ont été utilisés comme dans l'étude par Lee et Marsella[12] concernant la magnitude des mouvements de tête et les sourcils d'un orateur. Les participants de l'étude devaient noter leur sentiment immédiat d'un agent virtuel selon ces mouvements, et trois algorithmes d'apprentissage (Hidden Markov Model, Conditional Random Fields et Latent-Dynamic Conditional Random Fields) ont été comparés. Cependant, bien qu'ils aient amélioré la reconnaissance avec la dernière solution, la comparaison en terme de génération d'attitude par rapport aux résultats de la littérature n'a pas montré de différence significative.

Ravenet et al[13] ont créé un corpus de postures d'ECA selon différentes attitudes. Des utilisateurs devaient sélectionner une expression faciale et une amplitude de geste pour exprimer une attitude avec une intention conversationnelle (exprimer son accord avec une attitude soumise ou poser une question gentiment par exemple). Ils ont ainsi développé un modèle bayésien pour générer automatiquement des attitudes mais qui ne prend pas en compte la temporalité des signaux pour l'exprimer.

Enfin, l'utilisation d'algorithmes de sequence-mining a été explorée pour trouver des motifs utilisables en entrées d'algorithme de machine-learning pour faire de la génération d'agents. En particulier, Martinez et al.[14] puis Chollet et al.[15] expliquent comment les utiliser pour trouver des séquences simples de signaux non-verbaux associées à des attitudes sociales.

Martinez et al.[14] se placent dans le contexte des jeux vidéos pour relier des données à des émotions comme la frustration. Ils utilisent l'algorithme *Generalised Sequence Pattern* (GSP) sur des signaux physiologiques pour prédire l'état affectif du joueur. Cependant, ces séquences ne sont pas utilisées pour de la génération.

Chollet et al.[15] utilisent également GSP pour trouver des séquences de signaux sociaux annotés manuellement caractérisant différentes attitudes sociales. Néanmoins, GSP trouve des séquences d'événements sans l'information temporelle *i.e.* il ne peut trouver que l'ordre dans lequel les événements se produisent sans l'information sur le temps les séparant ou leurs durées. Ensuite, un modèle pour l'expression d'une attitude particulière par un ECA a été construit pour sélectionner la séquence la plus pertinente.

L'information temporelle reste donc la partie manquante de ces solutions alors qu'elle est importante car elle peut changer l'interprétation d'une séquence e.g. un long sourire opposé à un court comme montré dans [7].

3. NOTRE APPROCHE

Nous avons voulu continuer cette approche avec un algorithme de sequence-mining tout en prenant en compte cette information temporelle manquante. Pour cela, il faut d'abord transformer les signaux sociaux en événements temporels (voir 3.1 et 3.2 afin de pouvoir les utiliser en entrée de l'algorithme choisi : *Temporal Interval Tree Association Rule Learning* (TITARL) (voir 3.3).

3.1 Traitement des données

Dans un premier temps, pour validation, le set de signaux sociaux a été restreint aux *informations prosodiques*, aux *activations des AUs* et aux informations sur *les tours de parole*. D'autres données pourront être ajoutées comme le regard ou les mouvements de la tête. Les trois caractéristiques sélectionnées vont maintenant être décrites.

Les tours de paroles indiquent si l'humain est en train d'écouter ou de parler ainsi que les moments de prise et de fin de parole. Ces informations proviennent des transcriptions fournies par le corpus étudié (voir 4.1). Ces données sont aussi utilisées pour "décorer" d'autres événements comme les AUs en fonction de l'état : locuteur ou auditeur.

Les descripteurs prosodiques ont été extraits avec Prosogram, un programme développé par Mertens[16] qui fournit une représentation de l'intonation similaire à celle perçue par un humain. Prosogram a été préféré à d'autres outils d'extraction automatique de prosodie à cause de son approche phonétique qui ressemble plus à la perception humaine. En effet, tous les mouvements du pitch ne peuvent être perçus par l'oreille humaine et, comme le but à long terme est la génération, seul les signaux jouant un rôle dans la perception ont été considérés.
Par ailleurs, Prosogram propose une segmentation automatique des fichiers audio en syllabes notées nucleï puis calcule des paramètres prosodiques globaux tels que la gamme de pitch du locuteur. Il transforme ensuite cela en approximation des mouvements de pitch perçus. Cette approche "bottom-up" a l'avantage de ne pas avoir besoin d'informations supplémentaires (annotations, entraînements, ...) et limite donc le risque de biais.
Pour l'instant, les caractéristiques calculées sont, pour chaque nucleï, la fréquence fondamentale moyenne (f_0), ses variations, les pics d'intensité et la forme du pitch (montée, descente, plat, ...). Les formes du pitch ont ensuite été fusionnées grâce aux transcriptions fournies dans le corpus afin d'obtenir des descripteurs pour chaque mot prononcé. Les trois autres données, f_0, ses variations et les pics d'intensité, étant continues, elles sont transformées en événements symboliques avec le processus SAX expliqué dans 3.2.

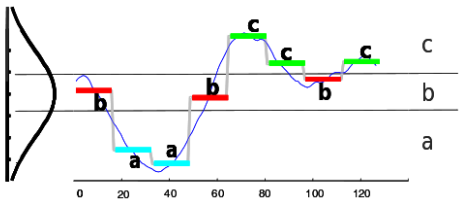

Figure 3: Représentation SAX d'un signal continu, image provenant de Lin et al.[18]

Les Action Units ont été automatiquement extraites grâce à la solution de Nicolle et al.[17] dont les résultats au challenge Fera 2015 nous ont convaincus de son efficacité. Afin de réduire le bruit de cette détection automatique, un lissage exponentiel (*"exponential smoothing"*) a été appliqué avec $\alpha = 0.5$ déterminé empiriquement. Les AUs sont ensuite symbolisées selon trois cas possibles : désactivée, faible activation et forte activation. Les études présentées ensuite se limitent aux AUs des sourcils (1 et 2 regroupées, pour le haussement, 4 pour le froncement), les pommettes (la 6) et les coins des lèvres (la 12) (voir fig.1).
Les variations d'activation de ces AUs sont considérées comme par exemple *AU6 de désactivée à faible* sera noté $AU6_{\text{off to low}}$ ou *AU12 de fort à faible* $AU12_{\text{high to low}}$. De plus, chaque événement est décoré selon l'état de la personne : locuteur ou auditeur.

3.2 Symbolisation en événements

Pour pouvoir utiliser l'algorithme de sequence-mining sur les signaux sociaux, ces derniers doivent être transformés en événements temporels symboliques. Afin d'être indépendants de l'utilisateur, les signaux sont centrés-normalisés, puis, le procédé de symbolisation se divise en deux cas selon le nombre de valeurs possibles pour le signal d'entrée.

Pour les signaux discrets ou avec une cardinalité faible comme les tours de paroles ou les AUs, la valeur réelle est utilisée (tour de parole) ou un léger regroupement est effectué (AUs).

Pour les signaux continus comme les descripteurs prosodiques (f_0, pitch, ...) ou l'inclinaison de la tête, la symbolisation est effectuée avec la représentation de données *Symbolic Aggregate Approximation* (SAX)[18] qui est très efficace pour les données temporelles. Cette représentation permet de réduire la dimension du signal tout en conservant l'ordonnancement de celui-ci. Cela permet la recherche de motifs dans le signal d'entrée tout en étant plus efficace que les autres représentations [18]. Après une *Piecewise Aggregate Approximation* du signal normalisé, SAX attribue un symbole à chaque segment selon une table de valeurs limites prédéfinies. Ces valeurs limites divisent l'étendue des données possibles en régions équiprobables en supposant une distribution normale (voir fig 3).

3.3 Algorithme de sequence-mining

Afin d'ajouter l'information temporelle dans la recherche de séquences de signaux sociaux, nous avons étudié la sous branche du sequence-mining dédiée au *Temporal Constrained Systems*. Après une étude des solutions possibles (Chronicle, Episode, ...), l'algorithme *Temporal Interval Tree Association Rules Learning* (TITARL)[19] a été retenu pour sa flexibilité et sa capacité à modéliser l'incertitude et l'inexactitude temporelle entre des événements temporels. En effet, en plus de trouver des relations entre des événements sous forme de règles (avant/après), il en ressort également des contraintes temporelles précises comme, par exemple, *"si il y a un événement D au temps t, il y aura un événement C à t+5"*. Cette approche d'apprentissage temporel permet de représenter des relations temporelles imprécises et non déterministes entre des signaux sociaux considérés comme des événements.

Un règle d'association temporelle donne des informations sur la relation reliant des événements symboliques avec un aspect temporel. Dans mon cas, il s'agit de signaux sociaux (AUs, prosodie, ...) considérés comme des événements discrets après l'étape de symbolisation présentée dans 3.2. Par exemple, avec les signaux montrés dans l'image 4, une règle pourrait être *"Si il y a une activation de l'AU 4 en parlant à l'instant t, alors il y aura une activation de l'AU 9 entre t+Δt et t+3Δt* , ce qui peut être symbolisé par la règle 1 :

$$\text{prise de parole} \xrightarrow[\textit{locuteur}]{\Delta t, 2\Delta t} AU4_{\text{off to low}} \xrightarrow{\Delta t, 3\Delta t} AU9_{\text{off to low}} \quad (1)$$

Δt représente ici un pas de temps lié aux données telles que les frames des vidéos par exemple. Une donnée importante d'une règle est donc sa distribution temporelle, ici Δt, 3Δt mais d'autres caractéristiques sont aussi calculées pour estimer l'intérêt de cette règle. Pour la règle r (définie dans 2)

$$A \xrightarrow{\Delta t_{min}, \Delta t_{max}} B \quad (2)$$

alors la confiance en cette règle est la probabilité de ses prédictions d'être vraies (voir 3a). Son support représente le pourcentage d'événements expliqués par celle-ci (voir 3b). Enfin, TITARL assure une bonne précision de ces règles c'est à dire une faible dispersion de la distribution des événements A vérifiant r (voir 3c).

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

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

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

Le pipeline de fonctionnement de TITARL est présenté dans l'image 5 et est divisé en deux grandes parties. L'entrée est un set d'événements temporels et une première étape va calculer des "règles simples" avec une structure comme dans

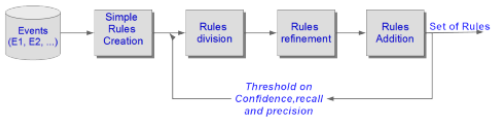

Figure 5: Pipeline de fonctionnement de TITARL

l'équation 2 qui auront une très grande distribution temporelle. Ces règles ont donc une très grande confiance et un support important mais une précision très mauvaise ce qui les rend non pertinentes. Par exemple, dans les habitats intelligents, si quelqu'un rentre dans une pièce, il en sera sorti après un temps infini. Cette règle est toujours vraie mais inintéressante car elle n'apporte pas d'information. Ce problème est cependant résolu avec les trois étapes suivantes qui seront appliquées récursivement : division, raffinement et addition.

L'étape de division fournit des règles plus pertinentes en gérant les cooccurrences d'événements. Par exemple, si il y a un événement A à l'instant t, un événement B à $t+5\Delta t$ et un autre B à $t+15\Delta t$, cela peut être symbolisé par la règle $A \xrightarrow{5\Delta t, 15\Delta t} B$ ou par les deux règles $A \xrightarrow{5\Delta t} B$ et $A \xrightarrow{15\Delta t} B$. L'étape de division permet de choisir entre ces deux propositions.

L'étape de raffinement a pour but d'augmenter la précision d'une règle. Il s'agit d'observer la distribution des événements vérifiant la règle et de réduire sa variance grâce à un seuillage sur son histogramme de distribution.

L'étape d'addition de conditions consiste à améliorer les règles obtenues en les fusionnant avec des règles simples divisées et raffinées. Pour cela, en prenant la règle $A \rightarrow B$, l'algorithme regarde les règles du type $B \rightarrow C$. Il les fusionne si le gain d'information est au-dessus d'un seuil fixé.

Plus de détails sur l'algorithme TITARL sont consultables dans l'article de Matthieu Guillame-Bert [19].

3.4 Évaluation et comparaison des règles d'associations

Afin de pouvoir comparer les règles entre elles, le score présenté dans l'équation 4 est utilisé. Il s'agit d'une combinaison pondérée de la confiance, du support et de l'intervalle temporel. Il a été introduit par Guillame-Bert dans [19] et testé sur un jeu de données artificielles ainsi qu'un corpus d'activités dans une résidence intelligente.

$$score = \frac{conf_r^4 . supp_r^2}{range_r} \quad (4)$$

Nous avons également défini la fréquence d'une règle comme

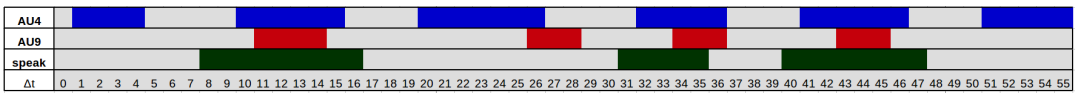

Figure 4: Exemple de timeline de signaux sociaux

étant le ratio entre son nombre d'occurrences et la durée des données. Cette donnée est une première étape afin de discriminer les règles liées à des raisons biomécaniques (mécanisme de la parole par exemple) des règles pertinentes pour l'étude des attitudes sociales sur des données réelles d'humains discutant entre eux.

4. APPLICATION À L'ÉTUDE DU CORPUS SEMAINE-DB

4.1 Semaine-Db

Pour illustrer notre méthode, elle a été utilisée sur la base de données SAL-SOLID SEMAINE [20]. Ce corpus utilise le paradigme *Sensitive Artificial Listener* (SAL) pour créer des interactions émotionnellement colorées entre un utilisateur et un 'caractère' joué par un opérateur. Il s'agit de flux vidéo et audio d'interactions dyadiques où l'opérateur répond avec des déclarations prédéfinies en fonction de l'état émotionnel de l'utilisateur.

Pour ces études, seule la partie opérateur a été considérée : à chaque session l'acteur joue quatre rôles prédéfinis correspondant aux quatre quadrants du circomplexe d'Argyle. Spike est agressif, Poppy est gentil, Obadiah est dépressif et Prudence pragmatique. Seuls les rôles de Poppy le gentil et Spike le méchant ont été retenus pour les comparer. Cela représente onze sessions d'enregistrements de 3-4 minutes comprenant 25 Poppy et 23 Spike joués par quatre acteurs différents.

Deux études ont été menées pour extraire des règles d'associations temporelles caractérisant l'attitude amicale et l'attitude hostile. Pour chaque étude, le but est de valider les règles obtenues en les comparant aux résultats vus dans la littérature. La première se concentre sur des sets d'AUs tandis que la seconde combine AUs et événements prosodiques.

4.2 Première étude : Action Units et attitudes sociales

Cette première étude met l'accent sur les AUs correspondant au sourire (AU6, AU12) et aux sourcils (AU1+2, AU4) afin de tester TITARL sur ces signaux sociaux spécifiques. En effet, nous avons voulu comparer les liens trouvés dans [21, 13] sur des études d'ECAs avec mes résultats. Ces articles soulignent qu'une attitude amicale comporte de nombreux sourires alors qu'une attitude hostile est exprimée par de nombreux froncements de sourcils.
Le tableau 1 montre des règles avec leurs confiances, supports, scores et ratios de fréquence. Il s'agit de règles avec l'un des meilleurs scores et un ratio de fréquence intéressant (i.e. discriminant). Ces résultats montrent que Poppy, le gentil, a plus tendance à sourire que Spike, l'hostile. Concernant la présence de l'AU 4 pour Spike, le faible ratio laisse à penser qu'il s'agit du mécanisme de production de la parole. Cela est renforcé par la présence de ces règles en mode locuteur.
En ce qui concerne les sourcils, il est confirmé que Spike les fronce beaucoup mais le résultat intéressant est sur le froncement de Poppy en mode auditeur. Cela peut être vu comme un backchannel indiquant l'intérêt de Poppy dans cette conversation au locuteur.

Ces résultats sont en accord avec la littérature et ajoutent à ceux-ci l'information temporelle et la confiance en ces règles. En effet, la recherche empirique et théorique ont montré qu'une attitude amicale implique des sourires fréquents alors que les froncements de sourcils sont liés à la menace et l'hostilité. Cette étude permet d'identifier de façon plus précise la durée de ces signaux sociaux. Cette information est très importante pour la génération d'une attitude par un ECA.

4.3 Seconde étude : variation de la voix, tour de parole, AUs et attitudes

Le but ici est de valider notre méthodologie avec des données audio et vidéo. Comme dans la première étude, les règles obtenues avec notre méthode sont donc comparées aux connexions à priori trouvées dans la littérature. Dans une étude du lien entre les sourcils et les variations de la voix lors des tours de parole, Guaïtelle et al. [22] montrent

	rule ($body \xrightarrow{\Delta t_{min};\Delta t_{max}} head$)	c	su	sc	fr
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: Exemples de règles trouvées par TITARL. La première partie montrent les liens trouvés entre les sourires et les mouvements de sourcils en fonction du personnage joué. La seconde présente les liens trouvés entre les mouvements de sourcils et la prosodie en fonction du personnage joué. Ces résultats sont présentés avec le rôle joué (Poppy/Spike) où ils sont le plus présent, leurs confiances (colonne c), leurs supports (su), leurs scores (sc) et leurs ratios de fréquence (rf).

que les pics du contour de la voix sont corrélés avec des mouvements similaires des sourcils (l'inverse étant faux cependant).

TITARL a permis d'étudier ces relations en calculant des règles d'associations entre les AU 1+2, AU 4, les variations de f_0 et la forme des mots. Les résultats montrés dans le tableau 1 soulignent le lien entre la forme de la f_0 au niveau du mot et la dynamique des sourcils : ces derniers suivant bien la forme du premier (un mot avec une f_0 *up and down* sera suivi par un haussement des sourcils). Cela peut être expliqué par l'expressivité des deux caractères joués et l'utilisation des sourcils pour exagérer l'accent tonique Anglais/Irlandais.
Cependant, aucune différence significative n'a été trouvée entre Poppy et Spike. Cela peut être dû à la métrique qui n'est pas adaptée à une différence aussi fine. Afin d'améliorer cela, des techniques venant du domaine de l'information retrieval vont être testées, en particulier sur la définition de mot clef pertinent.

Enfin, un résultat intéressant concerne la prise de parole de Poppy. Un haussement de sourcils de grande intensité suit souvent sa prise de parole et pourrait être utilisé pour améliorer le lien avec l'interlocuteur à ce moment précis et améliorer l'emphase.

5. CONCLUSION ET FUTURS DÉVELOPPEMENTS

Cet article présente les avancées pour mettre en place cette méthodologie pour extraire automatiquement des règles d'associations temporelles entre des signaux sociaux. Il se focalise principalement sur le traitement des signaux sociaux en entrée afin de permettre l'utilisation de l'algorithme de sequence mining voulu. Les premiers résultats valident cette démarche mais soulignent également les points à améliorer pour la suite.
Deux pistes sont explorées pour évaluer les règles calculées. La première consiste en une validation théorique basée sur un protocole *Leave-One-Subject-Out* où des règles obtenues sur un acteur seront comparées à celles obtenues sur les autres. La seconde repose sur la mise en place de la génération de l'avatar, également en fin d'implémentation, qui va permettre une évaluation subjective. Des observateurs vont évaluer des extraits vidéos, certains correspondant à l'occurrence d'une règle et d'autres à des extraits aléatoires, afin de valider leurs liens avec l'affect.

6. 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] K. R. Scherer. What are emotions ? and how can they be measured ? *Social science information*, 2005.

[6] Merijn Bruijnes, Rieks op den Akker, Sophie Spitters, Merijn Sanders, and Quihua Fu. The recognition of acted interpersonal stance in police interrogations and the influence of actor proficiency. *Journal on multimodal user interfaces*, 9(4) :353–376, 2015. Invited paper - Open Access.

[7] D. Keltner. Signs of appeasement : Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 1995.

[8] E. Bevacqua and C. Pelachaud. Expressive audio-visual speech. *Computer Animation and Virtual Worlds*, 2004.

[9] 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.

[10] 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.

[11] 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.

[12] J. Lee and S. Marsella. Modeling speaker behavior : A comparison of two approaches. In *IVA*, 2012.

[13] B. Ravenet, M. Ochs, and C. Pelachaud. From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In *IVA*, 2013.

[14] H. P. Martínez and G. N. Yannakakis. Mining multimodal sequential patterns : a case study on affect detection. In *ICMI*, 2011.

[15] M. Chollet, M. Ochs, and C. Pelachaud. From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In *IVA*, 2014.

[16] P. Mertens. The prosogram : Semi-automatic transcription of prosody based on a tonal perception model. In *Speech Prosody 2004, International Conference*, 2004.

[17] J. Nicolle, K. Bailly, and M. Chetouani. Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. *FERA*, 2015.

[18] J. Lin, E. Keogh, L. Wei, and S. Lonardi. Experiencing sax : A novel symbolic representation of time series. *Data Min. Knowl. Discov.*, 2007.

[19] M. Guillame-Bert and J. L. Crowley. Learning temporal association rules on symbolic time sequences. In *ACML*, 2012.

[20] 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.

[21] M. Ochs and C. Pelachaud. Model of the perception of smiling virtual character. In *AAMAS*, 2012.
[22] 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.