

---

# SMART : Règles d'associations temporelles de signaux sociaux pour la synthèse d'un Agent Conversationnel Animé avec une attitude spécifique

Thomas Janssoone<sup>1</sup>, Chloé Clavel<sup>2</sup>, Kévin Bailly<sup>1</sup>, Gaël Richard<sup>2</sup>

1. Pierre et Marie Curie University - ISIR

Paris, France

*prenom.nom@isir.upmc.fr*

2. LTCI, Télécom ParisTech, Université Paris Saclay

Paris, France

*prenom.nom@telecom-paristech.fr*

---

*RÉSUMÉ. Afin d'améliorer l'interaction entre des Humains et des Agents Conversationnels Animés (ACA), l'un des enjeux majeurs de ce domaine est de générer des agents crédibles socialement. Dans cet article, nous présentons une méthode capable de trouver des relations entre l'utilisation de signaux sociaux (mouvements de tête, expressions faciales, prosodie . . . ) par des Humains afin d'animer un ACA pour qu'il exprime une attitude sociale définie. Notre système, intitulé SMART pour Social Multimodal Association Rules with Timing, est basé sur un algorithme de sequence-mining qui lui permet de trouver des règles d'associations temporelles entre des signaux sociaux extraits automatiquement de flux audio-vidéo. Le calcul de ces règles est contrôlé pour faciliter ensuite l'animation d'un personnage virtuel. Dans ce papier, nous formalisons donc l'implémentation de SMART et nous justifions son intérêt par plusieurs études. Dans un premier temps, nous montrons que les règles calculées sont bien en accord avec la littérature en psychologie et sociologie. Ensuite, nous présentons les résultats d'évaluations perceptives que nous avons conduites suite à des études de corpus proposant l'expression d'attitudes sociales marquées.*

*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.*

*MOTS-CLÉS : Règles d'Association Temporelle, TITARL, Agents virtuels, attitudes sociales, traitement du signal social*

*KEYWORDS: Temporal Association Rules, TITARL, Virtual Agent, interpersonal stance, social signal processing*

## 1. Introduction

L'utilisation d'Agent Conversationnel Animé (ACA) offre une nouvelle façon d'interagir entre l'Humain et le machine. Ces ACA sont généralement des personnages virtuels qui font office d'interface entre l'utilisateur et un ou plusieurs programme. Ils peuvent par exemple aider des soldats lors du traitement d'un stress post-traumatique lié aux combats ou aider un patient à suivre son traitement (Truong *et al.*, 2015). L'un des principaux défis est donc de rendre cette interaction entre l'humain et l'ACA la plus fluide et naturelle possible. Une solution est de permettre à l'ACA d'exprimer différentes attitudes envers l'utilisateur, comme de la dominance pour un tuteur ou de la bienveillance pour un compagnon. Cet article va tout d'abord présenter notre méthode SMART, pour Social Multimodal Association Rules with Timing, dont une version préliminaire a été publiée dans Janssoone *et al.* (2016). Son but est la génération automatique de comportements réalistes pour un ACA avec une attitude spécifique grâce à l'analyse de l'utilisation de différents signaux sociaux (expressions faciales, mouvements de tête, prosodie, ...) pendant des interactions entre Humains afin d'en déduire des associations temporelles sous forme de règles. Nous présentons ici les derniers développements apportés et leurs évaluations afin de mettre en lumière les forces et limites de cette méthode. En particulier, nous présentons son application avec différentes échelles temporelles avec des ensembles cohérents de signaux sociaux, toujours en lien avec l'expression d'attitudes sociales.

Cette méthode s'inscrit donc dans le domaine du traitement du signal social qui est en pleine expansion (Vinciarelli *et al.*, 2009) et dont l'une de ses applications consiste à donner à des ACAs la capacité d'exprimer des émotions, des attitudes réalistes ou d'autres états émotionnels. En effet, comme Vinciarelli *et al.* (2012) l'explique, de nombreux corpus de travail et de nouvelles méthodes d'études ont été développées ces dernières années. Ces corpus, souvent composés de fichiers audiovisuels, fournissent les entrées pour des algorithmes d'apprentissages (Rudovic *et al.*, 2014; Pentland, 2004; Sandbach *et al.*, 2013; Savran *et al.*, 2014). Des humains experts ou différents algorithmes peuvent extraire des caractéristiques sur les signaux émis par le ou les protagonistes et les quantifier comme des descripteurs prosodiques (fréquence fondamentale de la voix, débit, intensité, ...) ou l'activation des muscles faciaux labellisés en Action Units (AUs, voir figure 1).

Les données sont généralement aussi annotées par un ou plusieurs observateurs extérieurs qui donnent ainsi leur perception de l'interaction en cours. Ces annotations fournissent alors différentes classes utiles pour les algorithmes d'apprentissage supervisés prenant en entrée les différents descripteurs.

Dans cet article, nous nous focalisons sur les attitudes sociales au sens de Scherer (2005) définies comme la "caractéristique d'un style affectif qui se développe spontanément 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

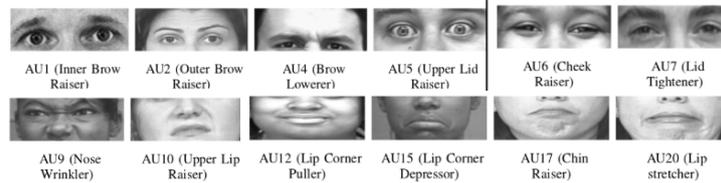


FIGURE 1. *Facial Action Unit correspondant à l'activation de différents muscles faciaux. Images obtenues via <http://www.cs.cmu.edu/~face/facs.htm>*

poli, distant, froid, chaleureux, compassionnel, dédaigneux) ". Les attitudes peuvent être estimées selon deux axes, l'un représentant l'appréciation et le second la dominance, permettant de définir le circomplexe interpersonnel, représentation proposée par Argyle (1975) et illustrée dans la figure 2.

Suivant les recommandations de Chindamo *et al.* (2012), nous nous concentrons sur l'étude de la dynamique de ces signaux car elle apporte de l'information sur l'attitude exprimée. En effet, la temporalité de certains signaux non-verbaux peut amener à différentes interprétations : Keltner (1995) illustre l'importance de cette dynamique avec l'exemple du sourire : un sourire long montre de l'amusement là où un regard fuyant suivi d'un sourire contrôlé peut signifier de l'embarras. Pour cela, nous proposons l'application de SMART à l'étude de la dynamique des signaux sociaux exprimés

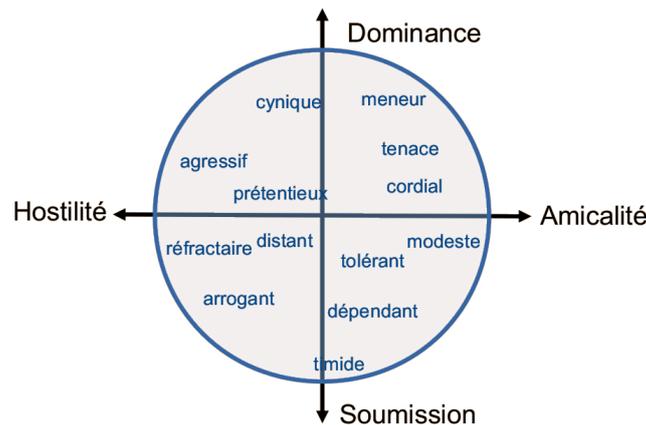


FIGURE 2. *Représentation du circomplexe interpersonnel, défini par Argyle*

par un protagoniste d'une interaction dyadique entre deux interlocuteurs.

## 2. Etat de l'art

La relation entre les signaux sociaux et les expressions sociales (émotions, attitudes, comportements, ...) a été étudiée durant les dernières décennies Vinciarelli *et al.* (2012). La principale méthode utilisée est de construire un modèle à partir d'observations d'humains exprimant ou réagissant à différents états émotionnels. Ce modèle est ensuite principalement utilisé pour deux cas d'application : soit pour de la détection sur un ou plusieurs sujets (e.g. détecter une émotion, du stress, de l'implication, ...), soit pour de la génération, par exemple en animant un ACA avec l'expression d'attitudes crédibles.

Nous allons maintenant passer en revue les principales méthodes employées, en commençant par des modèles psychologiques ou sociologiques basés principalement sur l'observation. Ils ont pu servir ensuite à la validation de modèles informatiques, appris sur des corpus. Ces derniers fournissent de nouveaux outils de recherche pour améliorer notre compréhension en intelligence artificielle, en interaction Homme-Machine ou en psychologie Marsella *et al.* (2010).

### 2.1. Méthodes orientées observation :

Plusieurs recherches utilisent l'observation d'interactions et en font des analyses qualitatives ou quantitatives afin de déterminer les signaux ayant influencés la perception des attitudes exprimées par les intervenants.

Par exemple, Tusing, Dillard (2000) dans leur étude sur la dominance cherchent les caractéristiques prosodiques qui influent sur la perception de la dominance des intervenants (i.e. sont ils dominants ou soumis). Pour cela, ils utilisent des extraits vidéos d'acteurs prononçant un message et des caractéristiques comme l'énergie, la fréquence fondamentale, l'intensité ou le débit en sont extraites. Leurs analyses permet de relier certaines de ces caractéristiques à différentes variations de perception de dominance. Ils montrent que l'énergie de la voix ainsi que ses variations influent positivement sur la perception de dominance. Ainsi, un message prononcé avec plus d'énergie sera perçu comme plus dominant. Ils montrent aussi que plus le débit est élevé, plus le message est perçu comme dominant. A contrario, cette analyse montre aussi qu'il n'existe pas de lien significatif pour certains signaux : un résultat notable est l'absence d'association entre le jugement de dominance et la fréquence fondamentale moyenne pour les locutrices femmes.

Dans des études qualitatives, Cafaro *et al.* (2012) 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 (sourire, regard et proximité). Ils insistent en particulier sur le fait que la distance physique entre l'observateur et l'agent n'a pas d'impact sur le jugement de gentillesse mais que le sourire influe principalement sur

cette dimension.

Cowie *et al.* (2010) 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.

Dans un but de détection, quelques travaux essaient de prendre en compte la temporalité, principalement en utilisant plusieurs fenêtres d'analyse dont les tailles en temps diffèrent. C'est le cas des travaux de Ward, A. (2016) qui propose d'observer sur différentes fenêtres temporelles les variations de prosodie entre un expert et un novice de jeux vidéos. Il y trouve des co-occurrences intéressantes avec les différentes phases du jeu et le rôle de chacun des joueurs.

De même, Audibert (2007) propose une étude de la modélisation des expressions prosodiques des affects avec un intérêt porté sur la temporalité. Pour cela, il analyse la perception de l'état émotionnel lié à des stimuli audio dont le contour prosodique est contrôlé. Ces stimuli ont été construits grâce à un acteur qui a été piégé lors d'une expérience en magicien d'Oz à exprimer certaines émotions tout en énonçant des mots (i.e. il est amené à prononcer des noms de couleurs monosyllabiques (jaune, rouge et vert) avec une certaine émotion (anxiété, déception, dégoût, ...)). Il utilise pour son expérience le principe de "gates" pour observer la temporalité : il s'agit de couper le stimuli original en des points prédéfinis et de le compléter avec du bruit blanc. Il observe en particulier que les émotions négatives dépendent plutôt de la qualité de voix et du débit tandis que les expressions de joie et de satisfaction sont plus liées au contours prosodique. Il montre aussi que les contours très amples de l'expression de la satisfaction permettent une reconnaissance précoce de celle-ci. Il reste cependant à poursuivre cette étude pour savoir si elle peut être généralisée aux attitudes et autres états émotionnels.

Enfin, Barbulescu *et al.* (2016) proposent une analyse discriminante linéaire afin de déterminer quelles caractéristiques audiovisuelles permettent de discriminer différentes attitudes dramatiques. Ils montrent ainsi qu'il vaut mieux se placer au niveau de la phrase pour avoir une meilleure reconnaissance qu'à un niveau sémantique plus bas comme la syllabe. Ils effectuent ensuite un test perceptif de la reconnaissance d'attitudes en animant un avatar sur un enregistrement de voix. Les mouvements de l'avatar sont contrôlés par un acteur. Ils montrent ainsi que la  $F_0$  est la caractéristique qui obtient les meilleurs jugements perceptifs lors de la synthèse d'attitudes.

Toujours dans un objectif de synthèse, Bawden *et al.* (2015) effectuent une analyse prosodique du corpus Semaine (McKeown *et al.* (2012) qui sera présenté dans la section 4.2). Ils explorent les relations entre la personnalité, le type d'acte de dialogue et des caractéristiques prosodiques. Ils fournissent des recommandations pour l'utilisation d'un système de synthèse vocale, en particulier pour l'animation d'ACA. Le corpus a été annoté manuellement pour les actes de dialogue (assertifs, directifs, expressifs, ...) et les caractéristiques prosodiques extraites avec le logiciel Praat ont été vérifiées manuellement. Cette étude montre l'influence de la personnalité sur le type d'actes de dialogues utilisés lors d'une interaction. L'analyse prosodique montre également une relation entre la personnalité et certaines caractéristiques : une personnalité agressive est associée avec les plus fortes intensités, des personnalités joyeuses ou pragmatiques auront, elles, le plus de variations de pitch. Les contours prosodiques ont également été corrélés avec les actes de dialogue et montrent l'importance d'une analyse plus fine de la taxonomie. Cependant, le lien entre ces contours et la personnalité n'a pas été étudié à ce niveau.

## 2.2. Méthode orienté Machine-Learning :

Par ailleurs, des algorithmes de machine-learning ont été utilisés comme dans les travaux de Lee, Marsella (2012) qui ont proposé une étude sur la construction d'un modèle de l'amplitude des mouvements de tête et des mouvements de sourcils d'un orateur. Trois algorithmes d'apprentissage (Hidden Markov Model, Conditional Random Fields et Latent-Dynamic Conditional Random Fields (LD-CRF)) ont été comparés en deux temps. Tout d'abord, ils testent leurs capacités de prédire ces mouvements : ils ont effectué une validation croisée en apprenant sur 70% de leur corpus et en testant sur les 30% restant. Ils montrent ainsi les bonnes performances du LD-CRF pour cette tâche. Une étude perceptive est ensuite discutée : les participants de l'étude devaient noter selon 16 dimensions leur sentiment de l'agent qu'ils avaient vu dans des vidéos générées selon un modèle de la littérature ou leur modèle basé Machine-Learning. Ces résultats étaient ensuite agrégés selon trois dimensions : l'impression de compétence, de sympathie et de pouvoir. Finalement, même si cette étude n'a pas montré de différence significative entre leur modèle et celui de la littérature, elle démontre la faisabilité et l'intérêt d'un modèle construit automatiquement. Les auteurs estiment que cette limitation peut être, soit à cause d'un nombre trop faible de données d'apprentissage pour analyser la complexité des comportements présents dans le corpus d'apprentissage, soit à cause des critères de l'étude qui n'étaient pas adaptés.

Ravenet *et al.* (2013) ont créé un corpus de postures d'ACA 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 alors développé un modèle bayésien pour générer automatiquement des attitudes mais qui ne prend pas en compte la temporalité des signaux pour l'exprimer.

### 2.3. *Un focus sur la temporalité et le Sequence-Mining :*

Enfin, une dernière solution pour faire de la génération d'agents consiste à rechercher des motifs ensuite utilisables en entrées d'algorithme de machine-learning. Pour cela, Martínez, Yannakakis (2011) puis Chollet *et al.* (2014) proposent de l'utilisation d'algorithmes de sequence-mining pour trouver des séquences simples de signaux non-verbaux associées à des attitudes sociales.

Martínez, Yannakakis (2011) 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.* (2014) utilisent également GSP pour trouver des séquences de signaux sociaux annotés manuellement caractérisant différentes attitudes sociales. Ils trouvent ainsi les séquences de signaux minimales pour exprimer une intention avec une attitude donnée. 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 réseau bayésien construit un modèle pour l'expression d'une attitude particulière par un ECA qui enrichi les séquences minimales en signaux pour mieux exprimer l'intention communicative. Ils ont ainsi montré grâce à des études perceptives que cette approche améliore bien l'expression d'attitudes par un agent virtuel.

Cette approche a été également explorée tout récemment par Dermouche, Pelachaud (2016) où l'algorithme de sequence-mining Apriori est modifié afin d'y ajouter une composante temporelle. Son algorithme, HCApriori, prend en entrée une base de séquences d'événements temporels et trouve les plus fréquentes. Une seconde analyse permet ensuite de trouver des liens entre les temps de début de ces séquences et leurs durées grâce à des opérations de clustering. Les résultats trouvés sont cohérents avec la littérature mais n'ont pas été soumis à une études perceptive en terme de synthèse de comportement d'agents.

### 2.4. *Positionnement :*

L'information temporelle reste donc souvent 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 Keltner (1995). Nous savons cependant quels signaux influent sur les attitudes comme cela est résumé dans le tableau 1.

Dans Janssoone *et al.* (2016), nous présentons SMART, détaillé dans la section 3, qui propose l'utilisation de règles d'associations temporelles modélisant les liens entre divers signaux lors de l'expression d'attitudes sociales. Nous avons alors proposé des

Modalité	Références	Influence la dominance	Influence l'appréciation
Prosodie	Tusing, Dillard (2000)	l'énergie de la voix, ses variations et le débit. F0 moyenne pour les hommes	-
	Vinciarelli <i>et al.</i> (2009)	pitch	silences
	Audibert (2007)	-	Contour prosodique semble liés aux expressions positives
	Barbulescu <i>et al.</i> (2016)	Fréquence fondamentale au niveau de la phrase	Fréquence fondamentale au niveau de la phrase
Mouvements de tête	Bawden <i>et al.</i> (2015)	-	Intensité, pitch, ses variations et son amplitude
	Cowie, Sawey (2011)	-	orientation du mouvement
	Ravenet <i>et al.</i> (2013)	Inclinaison de la tête vers le haut ou vers le bas	Inclinaison de la tête vers bas ou sur le coté (head shift - head tilt)
Expressions faciales	Vinciarelli <i>et al.</i> (2009)	Influence des AUs et références correspondantes	Influence des AUs et références correspondantes
	Ravenet <i>et al.</i> (2013)	expressions faciales négatives ou neutres	expressions faciales négatives ou positives
	Cafaro <i>et al.</i> (2012)	-	sourire

TABLE 1. Résumé de la littérature sur l'influence de différents signaux sociaux selon les différents axes du circomplexe interpersonnel d'Argyle permettant l'évaluation de la perception d'attitude sociale.

études qui permettaient de valider l'intérêt de cette solution. Par ailleurs, ce système est basé sur l'algorithme de sequence-mining TITARL de Guillaume-Bert, Crowley (2012), pour *Temporal Interval Tree Association Rules Learning*, dont le but est de trouver des associations temporelles entre des événements symboliques. Son intérêt est d'apporter, en plus du lien entre les signaux étudiés, une information temporelle sur les délais séparant ces différents événements. Initialement développé pour des applications de domotique ou de surveillance médicale, TITARL est également utilisé très récemment dans un but de détection de l'évolution du rapport de dominance lors d'une interaction par Zhao *et al.* (2016) où ils soulignent l'intérêt de ces règles pour améliorer le comportement d'un agent.

Dans la suite de ce papier, nous présentons et poursuivons l'élaboration de notre approche SMART. Nous proposons en particulier d'utiliser différentes échelles de temps pour le traitement de nos signaux sociaux en entrée. Pour cela, nous rappelons d'abord la structure de SMART ainsi que les différents signaux sociaux que nous avons à notre disposition. Ensuite, nous proposons trois cas d'études, toujours avec un but de génération. Dans un premier temps, nous étudions directement nos signaux multimodaux avec une échelle de temps "classique" en seconde.

Ensuite, nous nous concentrons sur les contours prosodiques des phrases en les liant aux attitudes exprimées. En effet, l'utilisation d'algorithmes de machine-learning Fernandez *et al.* (2014) en général et de sequence-mining en particulier (Laskowski *et al.*, 2008; Chen *et al.*, 2002) pour l'étude et la synthèse de contours prosodiques a déjà été proposée. Cependant, à notre connaissance, cette approche n'a jamais été proposée pour lier ces contours avec l'expression d'une attitude ce que nous proposons ici en adaptant l'échelle temporelle comme détaillé dans la partie 4.4.

### 3. SMART : trouver l'information temporelle liant les signaux sociaux

Nous présentons ici la structure de notre chaîne de traitement SMART dont les principales étapes sont visibles dans la figure 3. Son but est d'analyser des signaux sociaux, comme les mouvements de têtes, les Actions Units ou la prosodie issus de vidéos contenant des interactions, et d'en déduire des associations permettant la synthèse automatique du comportement d'un agent virtuel avec une attitude donnée. En sortie de ce système, des fichiers permettant une synthèse du comportement d'un ACA sont générés (voir 3.5). Un effort particulier, détaillé dans la partie 3.1, a été fourni pour que la représentation des signaux en événements symboliques permettent une transposition facile des règles d'association temporelle trouvées en information utile pour les fichiers de sortie.

#### 3.1. Extraction des signaux et symbolisation

Ce système prend en entrée des fichiers audio-vidéo et en extrait automatiquement les valeurs de différents signaux sociaux. Ces signaux sont ensuite transformés en événements temporels symboliques grâce à un partitionnement obtenu par un seuillage de la distribution des valeurs prises par les différents signaux. La structure de ces événements est choisie en fonction de la structure requise pour la synthèse (norme BML et SSML, voir 3.5). Ces transformations permettent de passer à l'étape suivante

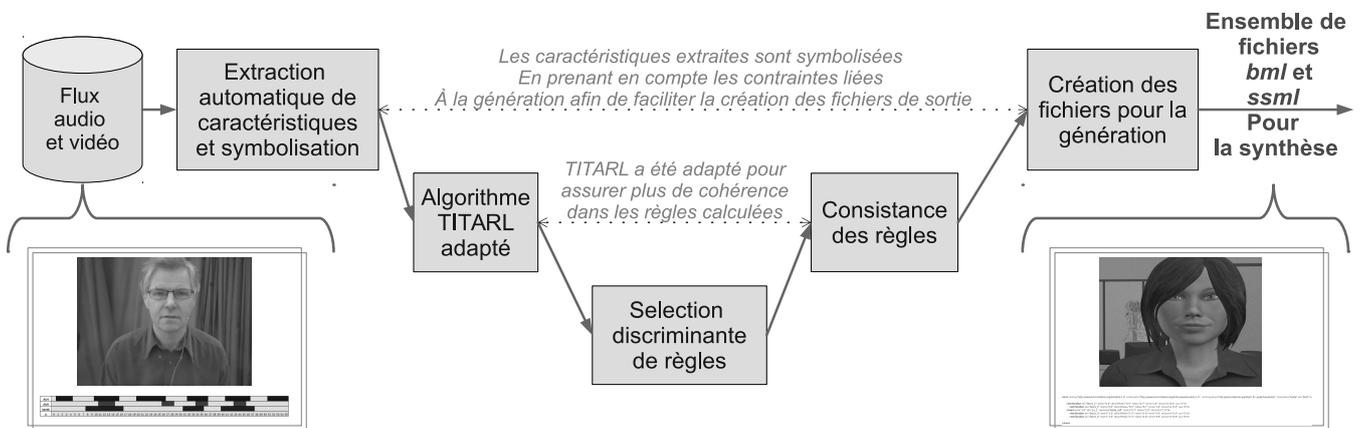


FIGURE 3. Schéma de fonctionnement de SMART. La ligne en pointillés souligne l'importance du type de fichier généré pour la synthèse et son impact sur l'étape de symbolisation des signaux sociaux.



TITARL assure en particulier une bonne précision de ces règles. Plus d'informations sur le fonctionnement de TITARL peut être trouvées dans le papier de Guillaume-Bert, Crowley (2012).

Les règles ainsi obtenues peuvent être ensuite complexifiées en ajoutant un événement en bout d'arbre. Une règle  $A \xrightarrow{t_1, t_2} B$  pourra être agrémentée de l'événement C pour obtenir la règle  $A \xrightarrow{t_1, t_2} B \xrightarrow{t_3, t_4} C$ . Cette étape pourra être répétée jusqu'à obtenir une taille voulue ou que les événements aux extrêmes de la règle correspondent à des événements voulus (début et fin de phrase par exemple).

### 3.3. Sélection des règles pertinentes

Le score indiqué dans l'équation 4 a été défini par Guillaume-Bert, Crowley (2012) comme une pondération entre la confiance, le support et l'intervalle temporel d'une règle. Il nous permet de classer les règles calculées en fonction de leur pertinence.

$$Score = \frac{conf_r^4 \cdot supp_r^2}{t_{max} - t_{min}} \quad (4)$$

Cependant, notre but est de relier ces règles d'associations à des attitudes données. Pour cela, une première adaptation de TITARL que nous proposons est de définir deux autres critères : la fréquence d'une règle pour une attitude donnée, équation 5a, et le ratio de fréquence, équation 5b. Ce dernier permet de discriminer si une règle est propre à une attitude ou est commune à plusieurs, voir toutes. Par exemple, si une règle apparaît souvent pour une attitude amicale et très peu pour une attitude hostile, elle peut être pertinente pour la synthèse d'un agent amical. A contrario, les règles correspondant aux mouvements de la mâchoire lors de la production de la parole ont des fréquences très proches pour toutes les attitudes et ne sont donc pas pertinentes.

$$Fréquence_{attitude_H}(r) = \frac{\text{occurrence d'une règle pour une attitude H}}{\text{durée des données pour l'attitude H}} \quad (5a)$$

$$\text{Ratio de fréquence}(r, attitude_H, attitude_F) = \frac{\text{fréquence}_{attitude_H}(r)}{\text{fréquence}_{attitude_F}(r)} \quad (5b)$$

### 3.4. Consistance des règles

La seconde adaptation proposée s'attaque au problème de cohérence des règles obtenues pour l'étape de génération. En effet, pour certains signaux comme les expressions faciales ou les mouvements de tête, les règles peuvent ignorer des événements importants pour assurer les transitions lors de la synthèse du comportement de l'ACA. En effet, les règles calculées par TITARL peuvent être de la forme présentée dans l'équation 6.

$$\text{Événement}_{\text{état 1 à état 2}} \xrightarrow{\Delta t_{min}; \Delta t_{max}} \text{Événement}_{\text{état 1 à état 2}} \quad (6)$$

Cette règle est intéressante pour de la détection mais, pour de la génération, il manque un événement pour que la règle soit cohérente : la transition de l'état 2 à l'état 1 est manquante. Par exemple, avec la règle *un haussement de sourcils est suivi 3 secondes plus tard par un haussement de sourcils*, la règle de synthèse doit intégrer des renseignements sur le moment de la baisse de ces sourcils.

Pour corriger ce problème, nous calculons à posteriori ces transitions en analysant les transitions possible et en forçant leur ajout dans l'arbre d'associations des règles. La règle de l'équation 6 devient alors celle présentée dans l'équation 7.

$$\text{Événement}_{\text{état 1 à état 2}} \xrightarrow{\Delta_{min_1} ; \Delta_{max_1}} \text{Événement}_{\text{état 2 à état 1}} \xrightarrow{\Delta_{min_2} ; \Delta_{max_2}} \text{Événement}_{\text{état 1 à état 2}} \quad (7)$$

Nous assurons ainsi la cohérence de la règle pour l'étape de génération en assurant que les événements correspondant à des changements d'états soient compatibles entre eux.

### 3.5. Transformation en BML et SSML

La dernière étape de notre système SMART consiste à transformer la règle d'association temporelle en fichiers pour la synthèse, *BML* et *SSML*. Le Behavior Markup Language (*BML*) est un langage de type XML qui permet le contrôle du comportement verbal et non-verbal d'un ACA. Un bloc *BML* décrit la réalisation physique de comportements (comme les expressions faciales, la parole, ...) et la synchronisation des contraintes entre ceux-ci. Le Speech Synthesis Markup Language, *SSML* est également basé sur le XML pour décrire les modifications de prosodie lors de la synthèse vocale de l'agent.

Grâce à ces fichiers, nous fournissons la temporalité de différents signaux sociaux exprimés par l'ACA pendant une animation. Pour cela, SMART retient, lors du calcul d'une règle, les occurrences de chaque événement la vérifiant et utilise comme temps de transition le  $\Delta_i$  ayant le plus d'occurrences (voir figure 5). Nous pouvons ainsi simplement trouver les temps de transitions nécessaire au *BML* et *SSML*. Dans une future version, nous pourrions utiliser les distributions des occurrences pour sélectionner différents temps de transitions et donc avoir beaucoup de variabilité dans la génération des animations.

## 4. Validation : études selon différents signaux sociaux et différentes échelles de temps

Dans cette partie, nous présentons les études que nous avons menées afin de montrer la validité de notre approche mais aussi ses limites. Dans un premier temps, nous détaillons les signaux sociaux qui seront étudiés et de quelle façon ils ont été extraits de fichiers audio-vidéo. Ensuite, le corpus étudié sera présenté et nous justifierons son choix avec notre problématique d'étude des attitudes sociales. Enfin, trois études se-

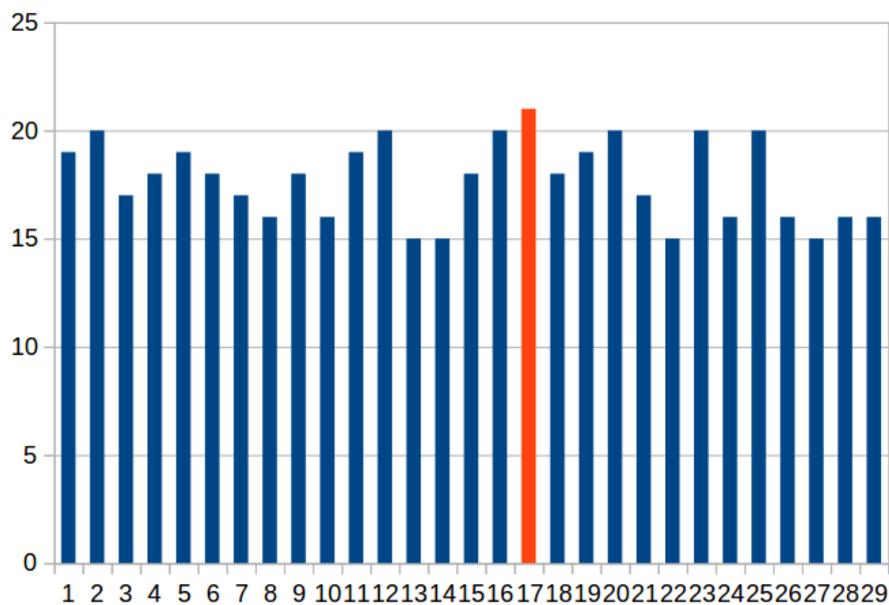


FIGURE 5. Exemple de distribution des occurrences des événements vérifiant une règle. Le  $\Delta_i$  ayant le plus d'occurrences est affiché en orange

ront détaillées : chacune traite différents signaux sociaux avec une échelle de temps spécifique. Différentes applications de SMART pour la synthèse d'attitudes sociales d'un ACA sont ainsi justifiées par rapport à la littérature et leurs résultats sont évalués.

#### 4.1. Les signaux étudiés

Dans un premier temps, pour validation, les set de signaux sociaux considérés ont été restreints aux *informations prosodiques*, aux *mouvements de tête*, aux *activations des AUs* et aux informations sur les *tours de parole*. D'autres données pourront être ajoutées comme le regard ou les gestes.

*Les tours de parole* 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 4.2). Ces données sont aussi utilisées pour qualifier d'autres événements comme les AUs en fonction de l'état, locuteur ou auditeur, et ajouter ainsi une information de contexte.

*Les descripteurs prosodiques* ont été extraits avec COVAREP (Degottex *et al.*, 2014) qui fournit un ensemble d'algorithmes de traitement de la parole afin de calculer plusieurs descripteurs. Dans cet article, nous nous limitons à l'utilisation de la fréquence fondamentale ( $F_0$ ) calculée toutes les 0.01 secondes pendant les tours de paroles (obtenus grâce aux transcriptions). L'étude 2 (cf 4.4) justifie cette limitation à cette seule caractéristique mais COVAREP présente également l'intérêt de proposer de nombreuses autres possibilités qui pourront être exploitées dans des études futures. Ce descripteur va être principalement utilisé pour générer des contours prosodiques relatifs comme cela est décrit dans la norme SSML<sup>1</sup>. Nous prenons donc en compte les contraintes liées à cette synthèse lors de notre étape de symbolisation et en particulier lors de la normalisation. En effet, pour chaque locuteur, nous calculons sa fréquence fondamentale moyenne pour chaque attitude étudiée, soit, en pratique, une  $F_0$  moyenne amicale et une  $F_0$  moyenne hostile, ainsi que leurs variances respectives. Puis, pour chaque  $F_0$  calculée par COVAREP, nous calculons le pourcentage d'écart par rapport à la  $F_0$  moyenne du locuteur avec l'attitude correspondante. Nous effectuons ensuite une partition des valeurs tous les 10 points de pourcentage. Cela facilitera l'étape de synthèse avec les SSMLplus détaillés dans la partie 4.4.

*Les Action Units* ont été automatiquement extraites grâce à la solution de Nicolle *et al.* (2015) dont les résultats au challenge Fera 2015 ont prouvé son efficacité. Elle permet de détecter les expressions faciales si un visage est présent et d'estimer leur intensité, de 0 (désactivé) à 5 (maximum). Afin de réduire le bruit de cette détection automatique, un lissage exponentiel ("*exponential smoothing*") a été appliqué avec  $\alpha = 0.5$  afin de supprimer les variations trop brutales. Les AUs sont ensuite symbolisées selon trois cas possibles : désactivée (valeur inférieure à 1), faible activation (valeur entre 1 et 3) et forte activation (valeur entre 3 et 5). Les études présentées ensuite se concentrent sur les AUs retenues par la littérature : celle des sourcils (1 et 2 regroupées, pour le haussement, 4 pour le froncement), des pommettes (la 6) et des 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 qualifié selon l'état de la personne : locuteur ou auditeur, comme cela a été indiqué dans la description *des tours de parole*.

## 4.2. Le corpus de travail

Afin d'illustrer et tester notre méthode, nous l'avons appliquée à la base de données SAL-SOLID SEMAINE (McKeown *et al.*, 2012). 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 prédéfi-

1. [https://www.w3.org/TR/speech-synthesis/#pitch\\_contour](https://www.w3.org/TR/speech-synthesis/#pitch_contour)

nies 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.

#### **4.3. Etudes 1 : Action Units, mouvements de tête et secondes**

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 Ochs, Pelachaud (2012); Ravenet *et al.* (2013) sur des études d'ACAs avec nos 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 2 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, l'amical, a plus tendance à sourire que Spike, l'hostile.

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 signal indiquant l'intérêt de Poppy dans cette conversation au locuteur. Enfin, nous retrouvons le lien entre les AUs et les mouvements de tête déjà précisé dans la littérature.

Ces résultats sont en accord avec la littérature et ajoutent à ceux-ci l'information temporelle et la confiance en ces règles. En effet, les recherches empiriques et théoriques 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.

Nous avons ensuite généré le comportement d'ACAs en fonction des meilleures règles trouvées par notre système, dont des exemples sont visibles ici<sup>2</sup>. Nous avons donc

---

2. <https://youtu.be/O2EPivej99Y>

mené une étude perceptive en demandant à des utilisateurs d'annoter l'attitude, hostile ou amicale, de l'agent dans des vidéos correspondant à ces règles. Pour cela, des annotateurs de la plate-forme *Crowdflower*<sup>3</sup> utilisaient une échelle de Likert en 5 points après avoir visionné la vidéo. Nous avons pu ainsi valider que les vidéos apprises sur des règles liées à Poppy était bien vues comme plus amicale que celle liées à Spike.

#### 4.4. Etude 2 : contours prosodiques, fréquence fondamentale et pourcentage

Dans cette étude, nous explorons une autre application de SMART dans le but de colorer la voix d'un ACA en fonction de l'attitude sociale planifiée. L'utilisation d'algorithme de data-mining, basés principalement sur du clustering, a déjà été exploré dans les domaines de la reconnaissance et de la synthèse vocale Laskowski *et al.* (2008); Chen *et al.* (2002), en particulier pour trouver des variations de fréquence fondamentale ( $F_0$ ) caractéristiques. Par ailleurs, la synthèse vocale d'un ACA, en particulier le contrôle de sa prosodie, peut se faire avec la norme *SSML*<sup>4</sup>. Nous utilisons donc les spécifications de cette norme pour orienter notre recherche : les contours prosodiques y sont définis comme une suite de doublets du type ( $x\%$ ,  $y\%$ ). Le premier élément,  $x$ , est un pourcentage temporel du texte contenu dans une balise *prosody* (voir figure 6). Dans notre cas, ce texte correspond à une phrase. Le second,  $y$ , est une valeur cible à atteindre pour la  $F_0$ . Dans notre application, nous prenons comme valeur cible  $y$ , un changement relatif exprimé en pourcentage par rapport à la fréquence fondamentale moyenne de l'interaction.

Un exemple de contrôle du contour prosodique avec la norme *SSML* est visible dans la

3. <https://www.crowdflower.com/>

4. <https://www.w3.org/TR/speech-synthesis/#S3.2.4>

	rule ( <i>body</i> $\xrightarrow{\Delta_{min};\Delta_{max}}$ <i>head</i> )	confiance	support	score	rf
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	$AU4_{\text{off to low /sp}} \xrightarrow{0.0s;0.9s} AU4_{\text{low to high /sp}} \xrightarrow{0.0s;0.7s} \text{head.yaw}_{[-10;10] \text{ to } [-20;-10]} \xrightarrow{0.0s;0.9s} AU4_{\text{low to high /sp}}$	0.38	0.02	$2.10^{-11}$	1.52
Poppy	$AU6_{\text{off to low /s}} \xrightarrow{0.1s;1.4s} \text{head.pitch}_{[-20;-10] \text{ to } [-30;-20]} \xrightarrow{1.2s;1.6s} AU6_{\text{off to low /s}} \xrightarrow{1.0s;1.8s} \text{head.pitch}_{[-30;-20] \text{ to } [-20;-10]}$	0.29	0.01	$2.10^{-12}$	1.64

TABLE 2. 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).

```

<?xml version="1.0" encoding="UTF-8" ?>
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.w3.org/2001/10/synthesis
  http://www.w3.org/TR/speech-synthesis/synthesis.xsd"
  xml:lang="en-US">

  <prosody pitch="199Hz" range = "53Hz"
    contour="(38%, -10%)(50%, +10%)(64%, +20%)">
    You might not be able to read this,
    but I do hope it reaches you somehow.
  </prosody>

</speak>

```

FIGURE 6. Exemple de contour prosodique défini avec la norme SSML

figure 6 où le paramétrage d'une phrase est détaillé. Son contenu verbal est "You might not be able to read this, but I do hope it reaches you somehow.". Elle sera prononcée à 199Hz de moyenne avec un écart possible de 53Hz, et son contour prosodique forcera une baisse de  $F_0$  de 10% à 38% de sa prononciation, une hausse de 10% à la moitié et, à 64% de son exécution, une hausse de 20%.

Nous avons donc repris les enregistrements des opérateurs de la base de données SAL-SOLID SEMAINE et, avec les descripteurs prosodiques décrits dans 4.1, nous créons un événement symbolique temporel avec comme valeur les variations de  $F_0$  et, en horodatage, le pourcentage du temps de cette variation par rapport au début et à la fin de la phrase. Nous utilisons donc ces événements temporels symboliques dans SMART et cherchons des règles d'associations temporelles dont le premier élément sera le début d'une phrase et, le dernier, la fin de cette même phrase. En plus de ces informations, chaque événement conserve l'information du genre du locuteur car l'état de l'art a montré une forte différence entre hommes et femmes. Afin d'améliorer nos résultats, nous avons effectué une validation-croisée pour évaluer les performances des règles sélectionnées dans une tâche de reconnaissance. Notre but était de nous assurer que les règles sélectionnées permettent une discrimination correcte et de trouver le seuil optimal à appliquer au ratio de fréquence. Nous avons obtenu une reconnaissance correcte avec un taux de validation de 75% pour un seuil de fréquence à 0.8. Nous obtenons ainsi des contours prosodiques au format de la norme SSML dont quelques exemples sont montrés dans le tableau 3.

Qualitativement, ces résultats sont en accord avec la littérature, en particulier l'étude de Bawden *et al.* (2015) qui portait sur le même corpus. En effet, les contours trouvés montrent généralement plus de variations de fréquence fondamentale chez Poppy que chez Spike car les règles trouvées comportent plus d'éléments, comme cela est vi-

	rule (body $\xrightarrow{\Delta_{min};\Delta_{max}}$ head)	contour	score
Poppy féminin	debut phrase $\xrightarrow{7\%;25\%} F_0(+50\%) \xrightarrow{51\%;69\%} F_0(+0\%) \xrightarrow{1\%;15\%} F_0(-10\%) \xrightarrow{3\%;17\%} F_0(+0\%) \xrightarrow{8\%;18\%} \text{fin phrase}$	contour="(15%,50%)(73%,0%)(80%,-10%)(90%,0%)"	$3,2 \cdot 10^{-16}$
Poppy féminin	debut phrase $\xrightarrow{13\%;19\%} F_0(-20\%) \xrightarrow{48\%;81\%} F_0(-30\%) \xrightarrow{0\%;16\%} F_0(-30\%) \xrightarrow{2\%;16\%} F_0(-20\%) \xrightarrow{0\%;16\%} \text{fin phrase}$	contour="(15%,-20%)(81%,-30%)(88%,-30%)(96%,-20%)"	$2,4 \cdot 10^{-18}$
Spike féminin	debut phrase $\xrightarrow{58\%;64\%} F_0(-20\%) \xrightarrow{0\%;22\%} F_0(-20\%) \xrightarrow{4\%;7\%} F_0(-20\%) \xrightarrow{5\%;21\%} F_0(-10\%) \xrightarrow{8\%;18\%} \text{fin phrase}$	contour="(60%,-20%)(69%,-20%)(74%,-20%)(87%,-10%)"	$2,9 \cdot 10^{-16}$
Spike féminin	debut phrase $\xrightarrow{55\%;67\%} F_0(+10\%) \xrightarrow{0\%;31\%} F_0(+10\%) \xrightarrow{4\%;7\%} F_0(+10\%) \xrightarrow{4\%;22\%} F_0(+20\%) \xrightarrow{2\%;12\%} \text{fin phrase}$	contour="(61%,10%)(75%,10%)(80%,10%)(93%,20%)"	$5,5 \cdot 10^{-17}$
Poppy masculin	debut phrase $\xrightarrow{5\%;99\%} F_0(-40\%) \xrightarrow{6\%;42\%} F_0(+10\%) \xrightarrow{15\%;27\%} F_0(-40\%) \xrightarrow{2\%;14\%} F_0(-30\%) \xrightarrow{20\%;30\%} \text{fin phrase}$	contour="(52%,-40%)(75%,10%)(90%,-40%)(95%,-30%)"	$5,5 \cdot 10^{-17}$
Poppy masculin	debut phrase $\xrightarrow{5\%;99\%} F_0(-40\%) \xrightarrow{47\%;54\%} F_0(-30\%) \xrightarrow{0\%;17\%} F_0(-30\%) \xrightarrow{0\%;5\%} F_0(-30\%) \xrightarrow{19\%;33\%} \text{fin phrase}$	contour="(22%,-40%)(60%,-30%)(75%,-30%)(79%,-30%)"	$1,7 \cdot 10^{-18}$
Spike masculin	debut phrase $\xrightarrow{0\%;6\%} F_0(-20\%) \xrightarrow{3\%;18\%} F_0(-10\%) \xrightarrow{68\%;86\%} F_0(-10\%) \xrightarrow{0\%;11\%} F_0(-10\%) \xrightarrow{0\%;10\%} \text{fin phrase}$	contour="(3%,-20%)(12%,-10%)(90%,-10%)(95%,-10%)"	$1,4 \cdot 10^{-16}$
Spike masculin	debut phrase $\xrightarrow{50\%;54\%} F_0(-20\%) \xrightarrow{1\%;27\%} F_0(-20\%) \xrightarrow{20\%;25\%} F_0(-20\%) \xrightarrow{0\%;16\%} F_0(-20\%) \xrightarrow{1\%;14\%} \text{fin phrase}$	contour="(51%,-20%)(64%,-20%)(86%,-20%)(93%,-20%)"	$4,1 \cdot 10^{-17}$

TABLE 3. Exemples de contours prosodiques de taille 4 trouvées avec la règle, le contour et le score selon le personnage joué. Nous présentons ici les deux meilleures règles trouvées pour Poppy et pour Spike pour chaque genre.

sible dans la figure 7. Un test de Mann-Whitney sur l'ensemble des règles pertinentes trouvées montrent également que celles trouvées pour Poppy ont une forte tendance à avoir plus d'associations que celles trouvées pour Spike, ( $p < 0.05$ ).

De plus, nous retrouvons que les valeurs des éléments composant les règles liées à Poppy sont généralement plus importante que chez Spike. Cela est visible dans le tableau 3 pour les règles de taille 4 (i.e. comportant 4 variations de  $F_0$ ). En effet, pour Poppy, l'écart possible moyen est de -17% avec une variance de 63%, tandis que pour Spike, cet écart est de -14% avec 31% de variance. On retrouve bien dans l'expression de l'amicalité plus de variance par rapport à de l'hostilité, comme l'avait souligné Audibert (2007).

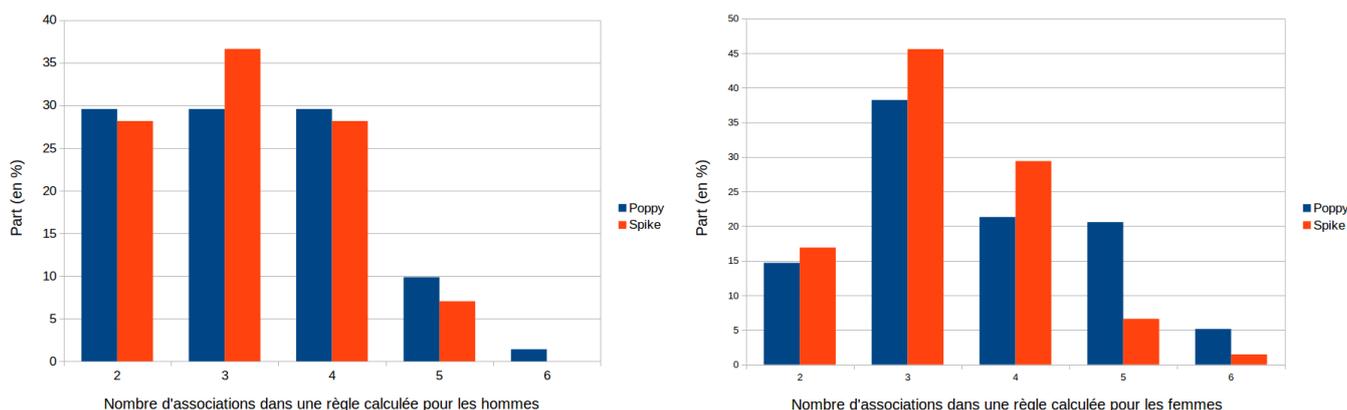


FIGURE 7. Nombre d'associations trouvées dans les règles selon le personnage joué

Afin de compléter notre étude, nous avons effectué une évaluation perceptive en générant grâce à un synthétiseur vocal des phrases prononcées avec les contours prosodiques correspondant aux meilleures règles. Pour cela, nous avons sélectionné dans les transcriptions originales de SEMAINE-DB deux phrases affirmatives et deux phrases interrogatives, pour chacune une prononcée par Spike et une par Poppy. Nous avons utilisé Mary (Modular Architecture for Research on speech Synthesis) TTS<sup>5</sup> comme synthétiseur vocal : il s'agit d'un logiciel libre en java qui est compatible avec la norme SSML. Il utilise la concaténation de diphones MBROLA, la sélection d'unités (choix pour un diphone du meilleur extrait d'enregistrement dans une base de sons) et des voix générées grâce à des modèles de Markov caché (MMC). Pour notre synthèse, nous utilisons les voix MMC appelées cmu-bdl-hsmm (masculine) et cmu-slt-hsmm (féminine) qui ont été construites à partir d'enregistrements faits à l'université de Carnegie Mellon.

Pour notre évaluation, nous prenons notre ensemble de phrase à évaluer et, pour chacune d'entre elles, nous les synthétisons avec les paramètres par défaut du synthétiseur afin d'obtenir des fichiers que nous appellerons neutre. Puis nous générons les fichiers audio avec la  $F_0$  moyenne calculée pour Poppy ou pour Spike sans toucher aux contours prosodiques ce qui nous donne deux références, nommées F0Spike et F0Poppy. Puis nous les synthétisons en prenant en compte aussi les contours prosodiques calculés par SMART et obtenons ContourSpike et ContourPoppy. Cela nous donne un ensemble de fichiers audio que nous faisons ensuite annoter en attitude sociale et en réalisme via la plate-forme Crowdfunder. Comme nous voulons analyser la synthèse vocale, les fichiers audio sont analysés et nous n'avons pas utilisé d'agents virtuels. Nous avons sélectionné les annotateurs résidant dans des pays anglophones et nous leur avons demandé via des échelles de Likert en 7 points de noter en amicalité et en réalisme les synthèses vocales de deux phrases, une affirmation et une interrogation, prononcée avec une voix féminine ou une voix masculine. Nous avons obtenu ainsi 283 jugements, visibles dans la figure 8.

Leur analyse montre tout d'abord que les phrases "brutes", sans modifications, ont été perçues comme légèrement amicale (75% des jugements supérieurs ou égal à 4). Ils montrent également que les modifications des différentes caractéristiques n'ont pas eu l'effet escompté, même si la modification de  $F_0$  rend Poppy plus sympathique. De même, le graphique montre que la modification du contour pour Spike a bien tendance à avoir un rendu plus hostile. Cependant, dans les deux cas, l'effet n'est pas significatif.

Pour comprendre cette absence d'effet, nous avons exploré les résultats. Une première piste est suggérée par Bawden *et al.* (2015) et consiste à regarder l'acte de dialogue. En effet, on voit que les tendances sont plus cohérentes avec nos attentes pour les affirmations que pour les questions, en particulier pour l'expression de l'hostilité. Ce-

---

5. <http://mary.dfki.de/>

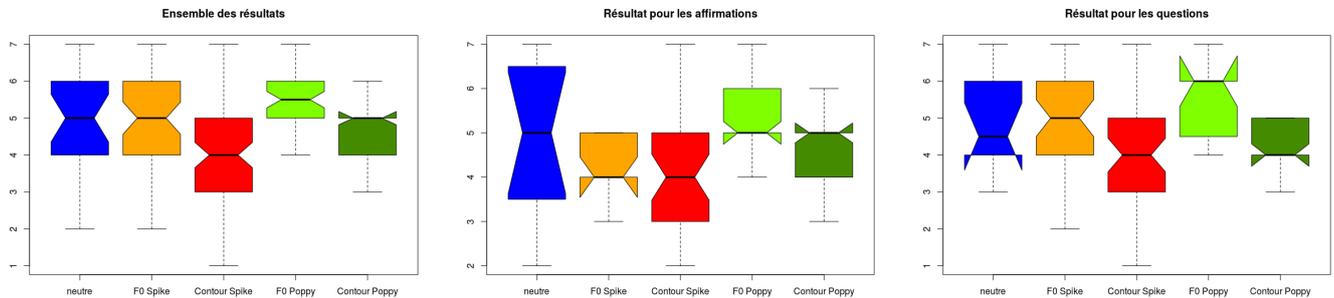


FIGURE 8. Graphiques représentant l'évaluation des fichiers en amicalité, de 1 très hostile à 7 très amical. En bleu, synthèse sans modification; En orange synthèse avec la  $F_0$  de Spike; En rouge synthèse avec les contours de Spike; En vert clair synthèse avec la  $F_0$  de Poppy; En vert foncé synthèse avec les contours de Poppy

pendant, ces résultats ne sont toujours pas significatif.

L'analyse des questions montre même que des données basées sur Spike sont vu plus amicale que le neutre. Nous remarquons aussi que nos résultats ont été perçus comme moins réalistes que les phrases neutre ou avec juste une modification de  $F_0$  globale. Effectivement, à l'écoute, la modification du contour peut avoir un rendu plus "robotique" ce qui peut être lié à la synthèse vocale basée sur un réseaux de Markov caché. En effet, ces méthodes paramétriques peuvent avoir un rendu non naturel qui est incompatible avec une modification subtile du contour prosodique. Cela se ressent d'ailleurs assez bien avec le jugement des annotateurs sur le réalisme de la voix. Par exemple, ces fichiers<sup>6</sup> ont été noté comme étant les plus réalistes. Il s'agit de synthèses où le contour a été modifié à partir de règles apprises sur Poppy et les annotateurs les ont d'ailleurs bien jugés amical. A contrario, ces synthèses<sup>7</sup>, basées également sur des contours appris sur Poppy, ont été jugées peu réaliste et hostile. La voix est très mécanique et rend donc l'écoute désagréable. Nous prévoyons donc de rechercher une solution efficace à ce problème de synthèse pour pouvoir mener à bien notre étude perceptive.

6. <https://www.youtube.com/watch?v=bjUUuyfBJms>  
<https://www.youtube.com/watch?v=NZqrh74wX-s>  
<https://www.youtube.com/watch?v=CJJhiYU6MVU>

7. <https://www.youtube.com/watch?v=CIxoofmH7s4>  
<https://www.youtube.com/watch?v=X2oJGTH78Uc>

## 5. Conclusion et ouvertures

Cet article présente une méthodologie pour extraire automatiquement des règles d'associations temporelles entre des signaux sociaux. Plusieurs points sont soulignés ici : 1) le traitement des signaux sociaux en entrée afin de permettre la synthèse des séquences de signaux apprises en fonction des normes requises pour l'animation d'agents virtuels, 2) la gestion de la dynamique temporelle dans les séquences. Nos premiers résultats valident cette démarche, en particulier sur les associations d'action units avec une échelle temporelle "classique" en secondes. Nous avons ici principalement exploré la possibilité d'adapter cette méthode à une échelle de temps différentes pour trouver des informations sur les contours prosodiques et les attitudes sociales. Nous retrouvons des informations en adéquation avec la littérature mais notre étude perceptive montre qu'il reste des améliorations à fournir pour obtenir une synthèse vocale réaliste et capable d'exprimer une attitude comme nous le souhaitons. Nous comptons améliorer notre système SMART afin de dépasser ces limitations, principalement en explorant des variantes pour la sélection de règles pertinentes (en exploitant par exemple l'utilisation de la validation croisée pour optimiser les critères de sélection comme initié en 4.4). Ensuite, nous envisageons d'évaluer différentes stratégies pour la synthèse multimodale des séquences de signaux correspondant aux attitudes sociales. Nous prévoyons ainsi de continuer nos travaux sur les règles associant action units, mouvements de tête et descripteurs prosodiques afin de proposer un système de génération multimodal d'attitudes sociales chez un agent conversationnel animé capable de mixer BML et SSML.

notre démarche de validation afin d'améliorer la sélection de règles pertinentes pour la synthèse mais également une application à la reconnaissance pourrait enrichir ce modèle. Ensuite, nous envisageons d'évaluer différentes stratégies pour lier les différentes modalités avec des outils de synthèse performants et adéquats pour atteindre notre but : améliorer la synthèse d'attitudes sociales réalistes exprimées par un agent conversationnel animé.

### Remerciements

*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.*

### Bibliographie

- Argyle M. (1975). *Bodily communication*. Methuen Publishing Company.
- Audibert N. (2007). Morphologie prosodique des expressions vocale des affects: quel timing pour le décodage de l'information émotionnelle. *Actes des VIIèmes RJC Parole, Paris*.
- Barbulescu A., Ronfard R., Bailly G. (2016). Characterization of audiovisual dramatic attitudes. In *Interspeech*.

- Bawden R., Clavel C., Landragin F. (2015). Towards the generation of dialogue acts in socio-affective ecas: a corpus-based prosodic analysis. *Language Resources and Evaluation*.
- Cafaro A., Vilhjálmsdóttir H. H., Bickmore T., Heylen D., Jóhannsdóttir K. R., Valgardsson G. S. (2012). First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Iva*.
- Chen Y., Gao W., Zhu T., Ling C. (2002). Learning prosodic patterns for mandarin speech synthesis. *Journal of Intelligent Information Systems*.
- Chindamo M., Allwood J., Ahlsen E. (2012). Some suggestions for the study of stance in communication. In *Passat and socialcom*.
- Chollet M., Ochs M., Pelachaud C. (2014). From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In *Iva*.
- Cowie R., Gunes H., McKeown G., Vaclau-Schneider L., Armstrong J., Douglas-Cowie E. (2010). The emotional and communicative significance of head nods and shakes in a naturalistic database. In *Lrec*.
- Cowie R., Sawey M. (2011). *Gtrace-general trace program from queen's, belfast*.
- Degottex G., Kane J., Drugman T., Raitio T., Scherer S. (2014). Covarep - a collaborative voice analysis repository for speech technologies. In *Icassp*.
- Dermouche S., Pelachaud C. (2016). Sequence-based multimodal behavior modeling for social agents. In *Icmi*.
- Fernandez R., Rendel A., Ramabhadran B., Hoory R. (2014). Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks. In *Interspeech*.
- Guillame-Bert M., Crowley J. L. (2012). Learning temporal association rules on symbolic time sequences. In *Acml*.
- Janssoone T., Clavel C., Bailly K., Richard G. (2016). Using temporal association rules for the synthesis of embodied conversational agent with a specific stance. In *Iva*.
- Keltner D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*.
- Laskowski K., Edlund J., Heldner M. (2008). Learning prosodic sequences using the fundamental frequency variation spectrum. In *Speech prosody*.
- Lee J., Marsella S. (2012). Modeling speaker behavior: A comparison of two approaches. In *Iva*.
- Marsella S., Gratch J., Petta P. (2010). Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*.
- Martínez H. P., Yannakakis G. N. (2011). Mining multimodal sequential patterns: a case study on affect detection. In *Icmi*.
- McKeown G., Valstar M., Cowie R., Pantic M., Schröder M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing*.

- Nicolle J., Bailly K., Chetouani M. (2015). Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. *FERA*.
- Ochs M., Pelachaud C. (2012). Model of the perception of smiling virtual character. In *Aamas*.
- Pentland A. (2004). Social dynamics: Signals and behavior. In *Icdl*.
- Ravenet B., Ochs M., Pelachaud C. (2013). From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In *Iva*.
- Rudovic O., Nicolaou M. A., Pavlovic V. (2014). 1 machine learning methods for social signal processing.
- Sandbach G., Zafeiriou S., Pantic M. (2013). Markov random field structures for facial action unit intensity estimation. In *Iccvw*.
- Savran A., Cao H., Nenkova A., Verma R. (2014). Temporal bayesian fusion for affect sensing: Combining video, audio, and lexical modalities.
- Scherer K. R. (2005). What are emotions? and how can they be measured? *Social science information*.
- Truong K., Heylen D., Chetouani M., Mutlu B., Salah A. A. (2015). Workshop on emotion representations and modelling for companion systems. In *Erm4ct@icmi*.
- Tusing K. J., Dillard J. P. (2000). The sounds of dominance. *Human Communication Research*.
- Vinciarelli A., Pantic M., Bourlard H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*.
- Vinciarelli A., Pantic M., Heylen D., Pelachaud C., Poggi I., D'Errico F. *et al.* (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *Affective Computing*.
- Ward N., A. S. (2016). Action-coordinating prosody. *Speech Prosody*.
- Zhao R., Sinha T., Black A., Cassell J. (2016). Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *Iva*.