Règles d'Associations Temporelles de signaux sociaux pour la synthèse d'Agents Conversationnels Animés :

Application aux attitudes sociales.

Thomas Janssoone [1], Chloé Clavel [2], Kévin Bailly [1], Gaël Richard [2]

1. *Sorbonne Universités, UPMC Univ Paris 06, CNRS, ISIR, 4 place Jussieu, 75252 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 du domaine est de générer des agents crédibles socialement. Dans cet article, nous présentons une méthode, intitulée SMART pour social multimodal association rules with timing, capable de trouver automatiquement des associations temporelles entre l'utilisation de signaux sociaux (mouvements de tête, expressions faciales, prosodie ...) issues de vidéos d'interactions d'humains exprimant différents états affectifs (comportement, attitude, émotions, ...). Notre système est basé sur un algorithme de fouille de séquences qui lui permet de trouver des règles d'associations temporelles entre des signaux sociaux extraits automatiquement de flux audio-vidéo. SMART va également analyser le lien de ces règles avec chaque état affectif pour ne conserver que celles qui sont pertinentes. Finalement, SMART va les enrichir afin d'assurer une animation facile d'un ACA pour qu'il exprime l'état voulu.*

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*

DOI:10.3166/RIA.31.511-537

1. Introduction

Le domaine du traitement du signal social est en pleine expansion (Vinciarelli *et al*., 2009) : il cherche, dans le même temps, à comprendre et modéliser les interactions sociales entre humains et à donner aux machines des capacités d'interactions similaires. Pour cela, comme Vinciarelli *et al*. (2012) l'expliquent, 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. Ces signaux peuvent provenir de l'audio comme les descripteurs prosodiques (fréquence fondamentale de la voix, débit, intensité, . . .) ou de la vidéo comme l'activation des muscles faciaux labellisés en Action Units (AUs, voir figure1). 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.

Ces progrès en traitement du signal conduisent à l'utilisation d'Agent Conversationnel Animé (ACA) comme interface avec un utilisateur où la machine a ainsi la capacité d'exprimer des émotions, des attitudes ou d'autres états affectifs. Ces ACA sont généralement des personnages virtuels qui font office d'interface entre l'utilisateur et un ou plusieurs programmes. 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. Le contrôle des "signaux sociaux" émis par l'ACA, comme ses expressions faciales ou sa synthèse vo-

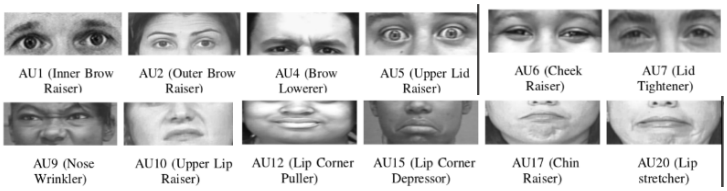

figure 1. Facial Action Unit correspondant à l'activation de différents muscles faciaux. Images obtenues via http://www.cs.cmu.edu/ ˜ face/facs.htm

cale, permet de lui faire exprimer différentes attitudes envers l'utilisateur, comme de la dominance pour un tuteur ou de la bienveillance pour un compagnon.

Cet article présente 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 tout en exprimant un état affectif spécifique. Pour cela, SMART analyse l'utilisation de différents signaux sociaux (expressions faciales, mouvements de tête, prosodie, tour de parole ...) extraits automatiquement de vidéos d'interactions entre humains afin d'en déduire des associations temporelles sous forme de règles qui sont liées à l'état affectif étudié (comportement, attitude, personnalité, ...). L'intérêt de SMART dans ce contexte est d'être particulièrement bien adapté pour transformer l'information contenue dans les règles d'associations temporelles entre les différentes variations de signaux sociaux en une information utile pour la synthèse de comportements d'ACA conforme aux normes en vigueur dans la communauté. 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 soulignons sa capacité d'adaptation à différentes mesures temporelles, seconde ou pourcentage d'élocution d'une phrase. Cet aspect est particulièrement intéressant pour l'expression d'attitudes sociales lors de la synthèse de comportement d'un ACA en restant cohérent avec les normes actuelles de la communauté [1].

Dans cet article, nous focalisons nos applications 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 (par exemple être 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, repré-

1. http://www.mindmakers.org/projects/saiba/wiki

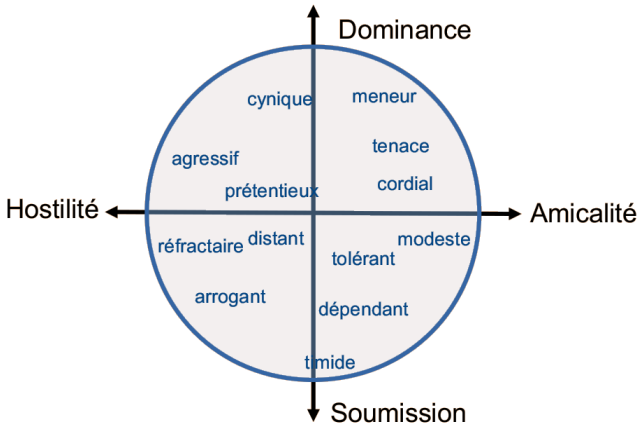

figure 2. Représentation du circomplexe interpersonnel, défini par Argyle

sentation 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 d'appliquer SMART à l'étude de la dynamique des signaux sociaux exprimés par un protagoniste d'une interaction dyadique entre deux interlocuteurs.

Dans la suite de ce papier, nous proposons une revue de l'état de l'art sur l'évolution de l'élaboration de modèles pour l'étude d'attitudes sociales. Nous présentons ensuite l'élaboration de notre approche SMART avec tout d'abord une introduction des règles d'associations temporelles et du formalisme inhérent. Nous soulignons ensuite nos contributions pour l'utilisation de ces règles pour la synthèse d'attitudes sociales d'un ACA.

Deux études de validations viennent illustrer notre approche sur un corpus audiovisuel dont différents signaux sociaux et états émotionnels ont été extraits. Chaque étude s'intéresse à différents types de signaux sociaux et va les étudier avec une échelle de temps adaptée à la synthèse d'un ACA.

Ainsi, dans la partie 4.3, nous étudions directement les AUs et mouvements de tête avec une échelle de temps "classique" en seconde. Nous retrouvons des résultats

conformes à la littérature qui justifient une étude perceptive pour évaluer ces résultats.

Ensuite, nous nous concentrons sur les contours prosodiques des phrases en les liant aux attitudes exprimées. En effet, l'utilisation d'algorithmes d'apprentissage automatique Fernandez *et al*. (2014) en général et de fouille de séquences 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 alors que des études, présentées dans la section 2, indiquent que cela est possible. Nous proposons donc de rechercher ce lien en adaptant l'échelle temporelle pour correspondre aux normes de synthèse vocale comme il est détaillé dans la partie 4.4.

Cela nous permet de conclure sur l'intérêt des règles d'associations temporelles pour enrichir la synthèse d'ACA, en particulier pour l'expression d'un état affectif voulu. Nous soulignons en particulier leur flexibilité pour la synthèse et l'information temporelle qu'elles apportent à partir de données faiblement annotées. Nous discutons des perspectives possibles de leurs applications pour des signaux multimodaux.

2. État 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 affectifs. 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.

En nous plaçant dans le contexte de synthèse d'ACA, nous allons maintenant passer en revue les principales méthodes employées en soulignant leurs évolutions et leurs limitations. Les premiers modèles psychologiques ou sociologiques sont basés principalement sur l'observation. L'outil statistique va ensuite permettre de trouver des informations complémentaires lors de l'analyse de ces observations et de descripteurs qui en ont été extraits. Ils ont pu servir ensuite à la validation de modèles informatiques, appris sur des corpus de données de plus en plus importants, qui vont pouvoir en extraire automatiquement des informations supplémentaires. Finalement, les algorithmes de fouille de séquences vont apporter une information temporelle essentielle concernant les modèles d'expressions d'attitudes sociales et améliorer donc la synthèse d'ACA.

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, dans leur étude sur la dominance, Tusing, Dillard (2000) 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 permettent 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 de durées différentes. 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-occurences 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. Il observe en particulier que les émotions négatives sont associées plutôt de la qualité de voix et du débit tandis que les expressions de joie et de satisfaction sont plus liées aux contours prosodiques. 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 la mieux corrélée aux jugements perceptifs lors de la synthèse d'attitudes pour l'ensemble des voix qu'ils ont étudiées.

Toujours dans un objectif de synthèse, Bawden *et al*. (2015) effectuent une analyse prosodique du corpus Semaine (McKeown *et al*. (2012), détaillé 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 (c'est-à-dire de variations aigu/grave). 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 orientées apprentissage automatique

Par ailleurs, des algorithmes d'apprentissage automatique ont été utilisés comme dans les travaux de Lee, Marsella (2012) qui ont étudié la construction d'un modèle de l'amplitude des mouvements de tête et des mouvements de sourcils d'un orateur. Trois algorithmes d'apprentissage (modèle de Markov caché, champs aléatoires conditionnels et *Latent-Dynamic Conditional Random Fields* (LD-CRF)) ont été comparés en deux temps. Tout d'abord, ils testent leurs capacités à prédire ces mouvements : ils ont effectué une validation croisée en apprenant sur 70 % du 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 le sentiment ressenti envers l'agent qu'ils avaient vu dans des vidéos générées selon un modèle de la littérature ou selon leur modèle basé sur de l'apprentissage automatique. 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 différences significatives 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 due au nombre trop faible de données d'apprentissage pour analyser la complexité des comportements présents dans le corpus d'apprentissage, soit due aux 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). Ravenet et al. ont alors développé un modèle bayésien pour générer automatiquement des attitudes mais ce modèle n'explore pas la temporalité pour moduler l'expression des attitudes sociales.

2.3. Un focus sur la temporalité et la fouille de séquences

Enfin, une dernière solution pour faire de la génération d'agents consiste à rechercher des motifs utilisables en entrée d'algorithmes d'apprentissage automatique. Pour cela, Martínez, Yannakakis (2011) puis Chollet *et al*. (2014) proposent d'utiliser des algorithmes de fouille de données 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), décrit dans Srikant, Agrawal (1996), 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 sur des signaux sociaux annotés manuellement afin d'en extraire des séquences 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 mais l'information temporelle reste limitée car il ne peut trouver que l'ordre dans lequel les événements se produisent sans l'information sur le temps les séparant ou leurs durées. Un réseau bayésien construit un modèle pour l'expression d'une attitude particulière par un ECA qui enrichit 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 d'exploration de données Apriori, également décrit dans Srikant, Agrawal (1996), est modifié afin d'y ajouter une composante temporelle. Son algorithme, HCApriori, va tout d'abord effectuer une opération de regroupement hiérarchique (*hierarchical clustering*) sur les signaux afin de trouver des liens entre l'instant du début de ces séquences et leurs durées ce qui donnera un ensemble de séquences temporelles. Celles ci seront analysées avec APriori pour trouver les motifs temporels fréquents et les lier à différentes attitudes sociales. Les résultats trouvés sont cohérents avec la littérature mais n'ont pas été soumis à une étude perceptive pour évaluer la synthèse de comportement d'agents.

2.4. Positionnement

Une information temporelle précise reste donc souvent la partie manquante de ces solutions pour générer efficacement des ACAs capable d'exprimer des attitudes sociales. Elle est d'autant plus importante qu'elle peut changer l'interprétation d'une séquence comme Keltner (1995) l'illustre avec l'exemple d'un long sourire opposé à un court. Les précédentes études montrent cependant quels signaux influent sur les attitudes comme cela est résumé dans le tableau 1.
Dans Janssoone *et al*. (2016), nous avons amorcé la mise en place de 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 avions ainsi mené des études qui permettaient de valider l'intérêt de cette solution. Par ailleurs, ce système est basé sur l'algorithme de fouille de séquences TITARL de Guillame-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 précise sur les délais séparant ces différents événements. Initialement développé pour des applications de prédiction en domotique ou en surveillance médicale, TITARL a également été utilisé très récemment dans un but de détection de l'évolution du rapport de dominance lors d'une interaction par Zhao *et al*. (2016) qui soulignent l'intérêt de ces règles pour améliorer le comportement d'un agent.
Nous proposons donc une utilisation de TITARL, encapsulée dans la méthode SMART, afin de pouvoir analyser des signaux bas niveaux, extraits automatiquement de flux audio-vidéos et donc potentiellement bruités. La fouille de séquences de TITATL permet d'en extraire des règles d'associations temporelles. Notre méthode va permettre de les lier à l'expression d'attitudes et va les enrichir pour permettre la synthèse du comportement d'un ACA exprimant l'attitude désirée.

tableau 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.

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 *et al.* (2009)	pitch	silences
	Audibert (2007)	-	Contour prosodique semble liés aux expressions positives
	Barbulescu *et al.* (2016)	Fréquence fondamentale au niveau de la phrase	Fréquence fondamentale au niveau de la phrase
	Bawden *et al.* (2015)	-	Intensité, pitch, ses variations et son amplitude
Mouvements de tête	Cowie, Sawey (2011)	-	orientation du mouvement
	Ravenet *et al.* (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 *et al.* (2009)	Influence des AUs et références correspondantes	Influence des AUs et références correspondantes
	Ravenet *et al.* (2013)	expressions faciales négatives ou neutres	expressions faciales négatives ou positives
	Cafaro *et al.* (2012)	-	sourire

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, qui ont été extraits automatiquement de vidéos contenant des interactions. SMART en déduit des associations permettant la synthèse automatique du comportement d'un agent virtuel avec une attitude donnée. Suivant les recommandations de l'état de l'art, l'information temporelle, particulièrement recherchée car c'est elle qui exprime le mieux l'attitude sociale, va être traduite sous forme de *règles d'association temporelle*, définies dans 3.2. Pour cela, un algorithme de fouille de séquence a été utilisé et adapté pour sélectionner les règles les plus pertinentes pour chaque attitude et pour répondre à notre deuxième contrainte : la synthèse de comportement pour un ACA. Ainsi, en sortie de ce système, des fichiers permettant une synthèse du comportement d'un ACA sont générés automatiquement (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.

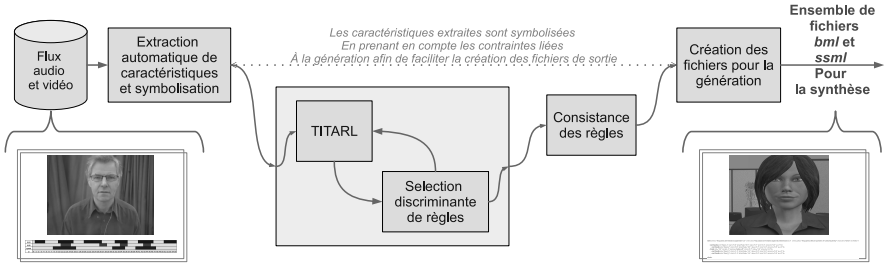

figure 3. Schéma de fonctionnement de SMART. La ligne en pointillés souligne l'importance du type de fichier généré pour la synthèse et son impact sur l'étape de symbolisation des signaux sociaux.

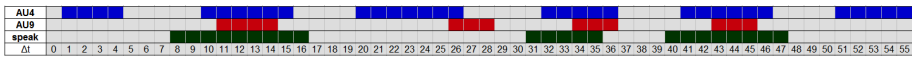

figure 4. Exemple de chronologie contenant les activations des AUs 4 et 9 ainsi que les tours de parole

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 qui est le cœur de notre système. Les différents outils utilisés pour cette extraction ainsi que les seuillages sont détaillés dans la partie 4.1.

3.2. TITARL et les règles d'associations temporelles

Cette étape consiste à analyser les événements grâce à un algorithme de fouille de séquence intitulé TITARL (Guillame-Bert, Crowley, 2012) afin d'extraire des règles d'association temporelle.

Une règle d'association temporelle donne des informations sur la relation entre les événements symboliques avec l'intervalle de temps les liant. Dans notre cas, il s'agit de signaux sociaux (AUs, prosodie, ...) considérés comme des événements discrets après l'étape de symbolisation. Tout d'abord, ils vont pouvoir donner une information contextuelle sur l'état de l'intervenant : locuteur ou orateur, homme ou femme, ... Un exemple de règle que l'on peut extraire des signaux montrés dans la figure 4 est : *"Une activation de l'AU 4 du locuteur à l'instant t sera suivi par une activation de l'AU 9 entre t+Δt et t+3Δt* , ce qui est formalisé par la règle 1 :

$$\underset{locuteur}{AU4_{\text{activation}}} \xrightarrow{\Delta t, 3\Delta t} \underset{locuteur}{AU9_{\text{activation}}} \tag{1}$$

Cela signifie que si le locuteur fronce des sourcils, il plissera du nez entre un Δt et 3Δt plus tard. 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 l'intervalle temporel qui la caractérise, ici [Δt, 3Δt] mais d'autres caractéristiques sont aussi calculées pour estimer l'intérêt de cette règle. Pour une règle de forme générale (2), notée r, traduisant que l'événement A sera suivi par l'événement B dans l'intervalle de temps $[t_{min}, t_{max}]$:

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

la confiance en cette règle est la probabilité, pour un événement A à l'instant t, d'avoir un événement B entre t+t_{min} et t+t_{max} (voir 3a). Son support représente le pourcentage d'événements expliqués par celle-ci (voir 3b). Enfin, la précision de ces règles est définie comme la dispersion de la distribution des événements A vérifiant r (voir 3c).

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

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

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

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 Guillame-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 extrémités 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 Guillame-Bert, Crowley (2012) comme une pondération entre la confiance, le support et l'intervalle temporel, défini par ses bornes t_{min} et t_{max}, d'une règle r. Il nous permet de classer les règles calculées en fonction de leurs pertinences.

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

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.

$$\text{Fréquence}_{\text{attitude}_H}(r) = \frac{\text{occurrence d'une règle pour une attitude H}}{\text{durée des données pour l'attitude H}} \tag{5a}$$

$$\text{Ratio de fréquence } (r, \text{attitude}_H, \text{attitude}_F) = \frac{\text{fréquence}_{\text{attitude}_H}(r)}{\text{fréquence}_{\text{attitude}_F}(r)} \tag{5b}$$

Nous prévoyons également d'étudier l'utilisation de ce ratio de fréquence pour évaluer différentes attitudes sociales selon les deux axes du "circomplexe" d'Argyle (voir fig 2). Par exemple, les attitudes *tenace*, *cordial* et *modeste* sont mesurées avec une amicalité similaire mais des degrés de dominance divers. Il serait intéressant de voir si les règles, communes à ces trois attitudes et absentes des autres, traduisent l'expression de cette amicalité.

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 la continuité des transitions lors de la synthèse du comportement de l'ACA. Cela peut être expliqué par la structure de TITARL pour le calcul des règles et par les signaux étudiés.

En effet, dans le calcul des règles, TITARL arrête son calcul lorsque les critères de qualité tels que la *confiance*, le *support* et la *precision* ne sont plus suffisant. TITARL a été initialement conçu pour des tâches de prédiction et c'est pourquoi il n'est pas adapté pour faire directement de la génération. Une de ces applications est la domotique et nous allons nous en inspirer pour illustrer ce problème : dans le cas de l'étude du comportement d'un utilisateur dans un appartement, on peut imaginer qu'une des règles trouvées soit *"quand l'utilisateur rentre dans la pièce, il allume la lumière dans la seconde qui suit"*, noté **R1** et formalisée dans 6a et qu'une seconde soit *"lorsque l'utilisateur éteint la lumière, il sort dans la seconde qui suit"*, noté **R2** et formalisée dans 6b

$$\textbf{R1 :}\ \text{Utilisateur}_{\text{entre}} \xrightarrow{0s;1s} \text{Utilisateur} \tag{6a}$$

$$\textbf{R2 :}\ \text{Utilisateur}_{\text{éteint la lumière}} \xrightarrow{0s;1s} \text{Utilisateur}_{\text{sort}} \tag{6b}$$

$$\textbf{R3 :}\ \text{Utilisateur}_{\text{allume la lumière}} \xrightarrow{0s;\Delta t} \text{Utilisateur}_{\text{éteint la lumière}} \tag{6c}$$

Le problème est que l'utilisateur peut passer un temps très variable dans la pièce et donc que la règle **R3**, visible dans 6c qui permet de relier l'allumage de la lumière avec son extinction va avoir un intervalle temporel très grand. La précision de R3 sera donc très mauvaise et son score également ce qui fait que cette règle ne sera pas retenue.

De plus, nos signaux ne sont pas à valeurs binaires ce qui augmente le risque de calcul de règles inconsistante. Ainsi, si on considère un signal S qui peut avoir trois valeurs notées v_1, v_2 et v_3, les règles calculées par TITARL peuvent être de la forme présentée dans (7).

$$\text{S}_{v_1 \text{ à } v_2} \xrightarrow{\Delta t_{min};\Delta t_{max}} \text{S}_{v_1 \text{ à } v_2} \tag{7}$$

Cette règle est intéressante pour de la détection mais, pour de la génération, il manque l'information sur la transition de l'état v_2 à v_1. Lorsque le signal S passe de v_1 à v_2, une partie peut ensuite retourner directement à v_1 tandis qu'une autre passera par v_3

avant de revenir à v_1. Cela aura une incidence sur le support et la confiance et empêcher le calcul automatique de cette transition. Ce problème est d'autant plus présent dans notre cas que les informations sur les signaux sociaux étudiés ont été extraites automatiquement et sont susceptibles d'être bruitées. Par exemple, avec la règle AU_1 *(haussement de sourcils) de désactivée à faible est suivi 3 secondes plus tard par* AU_1 *de désactivée à faible*, la règle de synthèse doit intégrer des renseignements sur le moment de la désactivation de l'AU_1. L'algorithme de détections des AUs peut avoir trouvé des activations fortes et la continuité dans la transition ne peut pas être trouvée par TITARL.

Pour corriger ce problème, nous calculons a posteriori ces transitions en analysant les transitions possibles et en forçant leur ajout dans l'arbre d'associations des règles. La règle de l'équation 7 devient alors celle présentée dans l'équation 8.

$$S_{v_1 \text{ à } v_2} \xrightarrow{\Delta t_{min_1};\Delta t_{max_1}} S_{v_2 \text{ à } v_1} \xrightarrow{\Delta t_{min_2};\Delta t_{max_2}} S_{v_1 \text{ à } v_2} \tag{8}$$

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 entres 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, lors du calcul d'une règle, SMART retient l'ensemble des suites d'événements la vérifiant et utilise comme temps de transition les Δ_t ayant le plus d'occurrences (voir figure 5). Nous pouvons ainsi trouver simplement les temps de transitions nécessaire au *BML* et *SSML*. Dans une future version, nous pourrons utiliser les distributions des occurrences pour sélectionner différents temps de transitions et introduire ainsi plus de variabilité dans la génération des animations. Cela est d'autant plus vrai que d'autres temps de transitions non retenu peuvent avoir un nombre d'occurrences très proche du maximum.

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

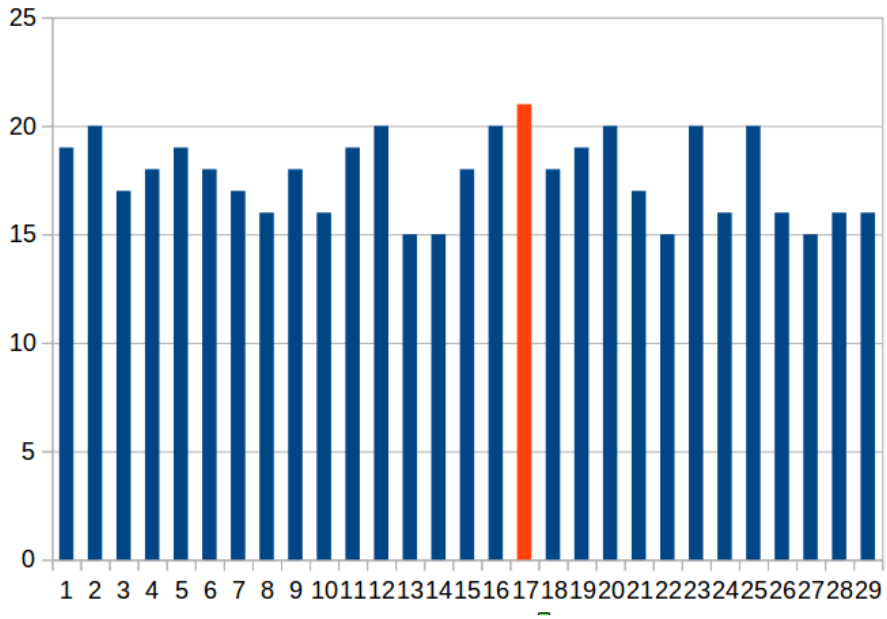

figure 5. Exemple de distribution des occurrences des événements vérifiant une règle. Le Δ_t ayant le plus d'occurrences est affiché en orange. A noté que ce choix du plus occurrent est perfectible car différents Δ_t pourraient correspondre ou être très proches de plus occurrent

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 par notre problématique d'étude des attitudes sociales. Enfin, deux études seront 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, l'ensemble des signaux sociaux considérés a été restreint 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 trans-

criptions fournies dans 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*[2]. 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 fichiers à la norme *SSML* détaillée 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 l'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 $\alpha = 0.5$ qui permet 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, 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*.

2. https://www.w3.org/TR/speech-synthesis/#pitch_contour

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 humain mais qui se comporte comme un agent. 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 "ircomplexe" 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. Un exemple de ces interactions est visible ici [3] où l'on voit la transition de l'acteur entre le rôle de Poppy avec celui de Spike.

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 ensembles 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 mouvements de 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, 1.25 dans notre cas). 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

3. https://youtu.be/Tt8UOw4-Mdw

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 : il fronce les sourcils pour exprimer sa concentration sur le discours de l'utilisateur. 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 [4]. Nous avons donc 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* [5] 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 amicales 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 fouille de données, qui s'appuie principalement sur *une étape de clustering*, a déjà été explorée dans les domaines de la reconnaissance et de la synthèse

4. https://youtu.be/O2EPivej99Y
5. https://www.crowdflower.com/

tableau 2. Exemples de règles trouvées par TITARL. La première partie montrent les liens trouvés entre les sourires (AU6 et AU12) et les mouvements de sourcils (AU4) en fonction du personnage joué. La seconde présente les liens trouvés entre les mouvements de sourcils et les mouvements de tête (pitch et yaw) 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)

	rule ($body \xrightarrow{\Delta t_{min};\Delta t_{max}} head$)	**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	$AU4_{\text{off to low /sp}} \xrightarrow{0.0s;0.9s} AU4_{\text{low to high /sp}} \xrightarrow{0.0s;0.7s} head.yaw_{[-10;10] \text{ to } [-20;-10]} \xrightarrow{0.0s;0.9s} AU4_{\text{low to high /sp}}$	0.38	0.02	2.10^{-11}	1.52
Poppy	$AU6_{\text{off to low /ls}} \xrightarrow{0.1s;1.4s} head.pitch_{[-20;-10] \text{ to } [-30;-20]} \xrightarrow{1.2s;1.6s} AU6_{\text{off to low /ls}} \xrightarrow{1.0s;1.8s} head.pitch_{[-30;-20] \text{ to } [-20;-10]}$	0.29	0.01	2.10^{-12}	1.64

```
<?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

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*[6]. 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 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 53 Hz, 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

6. https://www.w3.org/TR/speech-synthesis/#S3.2.4

tableau 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

	rule (*body* $\xrightarrow{\Delta t_{min};\Delta t_{max};}$ *head*)	**contour**	**score**
Poppy féminin	debut phrase $\xrightarrow{7\%;25\%;}$ $F_0(+50\%)$ $\xrightarrow{51\%;69\%;}$ $F_0(+0\%)$ $\xrightarrow{1\%;15\%;}$ $F_0(-10\%)$ $\xrightarrow{3\%;17\%;}$ $F_0(+0\%)$ $\xrightarrow{8\%;18\%;}$ fin phrase	contour="(15%,50%)(73%,0%)(80%,-10%)(90%,0%)"	$3,2.10^{-16}$
Poppy féminin	debut phrase $\xrightarrow{13\%;19\%;}$ $F_0(-20\%)$ $\xrightarrow{48\%;81\%;}$ $F_0(-30\%)$ $\xrightarrow{0\%;16\%;}$ $F_0(-30\%)$ $\xrightarrow{2\%;16\%;}$ $F_0(-20\%)$ $\xrightarrow{0\%;16\%;}$ fin phrase	contour="(15%,-20%)(81%,-30%)(88%,-30%)(96%,-20%)"	$2,4.10^{-18}$
Spike féminin	debut phrase $\xrightarrow{58\%;64\%;}$ $F_0(-20\%)$ $\xrightarrow{0\%;22\%;}$ $F_0(-20\%)$ $\xrightarrow{4\%;7\%;}$ $F_0(-20\%)$ $\xrightarrow{5\%;21\%;}$ $F_0(-10\%)$ $\xrightarrow{8\%;18\%;}$ fin phrase	contour="(60%, -20%)(69%, -20%)(74%, -20%)(87%, -10%)"	$2,9.10^{-16}$
Spike féminin	debut phrase $\xrightarrow{55\%;67\%;}$ $F_0(+10\%)$ $\xrightarrow{0\%;31\%;}$ $F_0(+10\%)$ $\xrightarrow{4\%;7\%;}$ $F_0(+10\%)$ $\xrightarrow{4\%;22\%;}$ $F_0(+20\%)$ $\xrightarrow{2\%;12\%;}$ fin phrase	contour="(61%, 10%)(75%, 10%)(80%, 10%)(93%, 20%)"	$5,5.10^{-17}$
Poppy masculin	debut phrase $\xrightarrow{5\%;99\%;}$ $F_0(-40\%)$ $\xrightarrow{6\%;42\%;}$ $F_0(+10\%)$ $\xrightarrow{15\%;27\%;}$ $F_0(-40\%)$ $\xrightarrow{2\%;14\%;}$ $F_0(-30\%)$ $\xrightarrow{20\%;30\%;}$ fin phrase	contour="(52%, -40%)(75%, 10%)(90%, -40%)(95%, -30%)"	$5,5.10^{-17}$
Poppy masculin	debut phrase $\xrightarrow{5\%;99\%;}$ $F_0(-40\%)$ $\xrightarrow{47\%;54\%;}$ $F_0(-30\%)$ $\xrightarrow{0\%;17\%;}$ $F_0(-30\%)$ $\xrightarrow{0\%;5\%;}$ $F_0(-30\%)$ $\xrightarrow{19\%;33\%;}$ fin phrase	contour="(22%, -40%)(60%, -30%)(75%, -30%)(79%, -30%)"	$1,7.10^{-18}$
Spike masculin	debut phrase $\xrightarrow{0\%;6\%;}$ $F_0(-20\%)$ $\xrightarrow{3\%;18\%;}$ $F_0(-10\%)$ $\xrightarrow{68\%;86\%;}$ $F_0(-10\%)$ $\xrightarrow{0\%;11\%;}$ $F_0(-10\%)$ $\xrightarrow{0\%;10\%;}$ fin phrase	contour="(3%, -20%)(12%, -10%)(90%, -10%)(95%, -10%)"	$1,4.10^{-16}$
Spike masculin	debut phrase $\xrightarrow{50\%;54\%;}$ $F_0(-20\%)$ $\xrightarrow{1\%;27\%;}$ $F_0(-20\%)$ $\xrightarrow{20\%;25\%;}$ $F_0(-20\%)$ $\xrightarrow{0\%;16\%;}$ $F_0(-20\%)$ $\xrightarrow{1\%;14\%;}$ fin phrase	contour="(51%, -20%)(64%, -20%)(86%, -20%)(93%, -20%)"	$4,1.10^{-17}$

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 visible 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).

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 [7] comme synthétiseur vocal : il s'agit d'un logiciel libre en java qui est compatible avec la norme

7. http://mary.dfki.de/

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 fait à l'université de Carnegie Mellon.

Pour notre évaluation, nous prenons notre ensemble de phrases 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*.

Enfin 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 Crowdflower. 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).

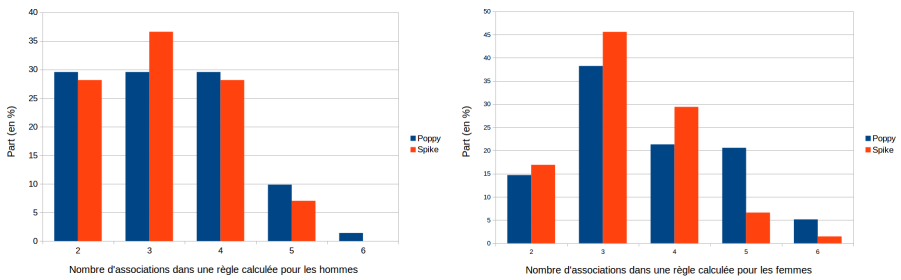

figure 7. Nombre d'associations trouvées dans les règles selon le personnage joué

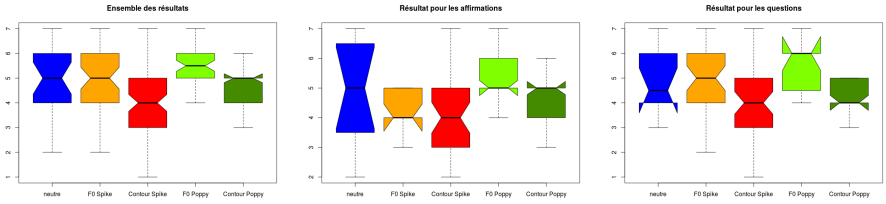

figure 8. Graphiques représentant l'évaluation des fichiers en amicalité, de 1 très hostile à 7 très amical. En bleu, synthèse sans modification ; En orange, synthèse avec la F_0 de Spike; En rouge, synthèse avec les contours de Spike; En vert clair, synthèse avec la F_0 de Poppy; En vert foncé, synthèse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

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 a 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. Cependant, ces résultats ne sont toujours pas significatifs.

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 neutres 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 réseau 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 [8] 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

8. https://www.youtube.com/watch?v=bjUUuyfBJms
https://www.youtube.com/watch?v=NZqrh74wX-s
https://www.youtube.com/watch?v=CJJhiYU6MVU
—

ont d'ailleurs bien jugés comme amicaux. A contrario, ces synthèses [9], 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.

Pour éviter cette limitation, nous avons choisi d'évaluer des fichiers audio de voix dont les contours prosodiques ont été modifiés précisément grâce à Promo [10]. Il s'agit d'une librairie qui permet de manipuler la fréquence fondamentale et qui fait de la re-synthèse en se basant sur Praat (Boersma, Weenink, 2017). Nous pouvons ainsi évaluer l'impact des contours prosodiques dans la perception de l'attitude sociale. De la même façon, nous avons demandé à 120 utilisateurs anglophones de a plateforme Crowdflower d'évaluer leur perception de l'attitude sociale exprimées dans chaque fichier audio. Chaque sujet en avait 4 à évaluer : un audio avec un contour plat et une F_0 amicale (i.e. la F_0 moyenne calculée sur les sessions Poppy), un audio avec un contour plat hostile (i.e. la F_0 moyenne calculée sur les sessions Spike), un audio avec un contour prosodique amical (F_0 moyenne et contour prosodique calculé sur Poppy)

9. https://www.youtube.com/watch?v=CIxoofmH7s4
https://www.youtube.com/watch?v=X2oJGTH78Uc
—
10. https://github.com/timmahrt/ProMo

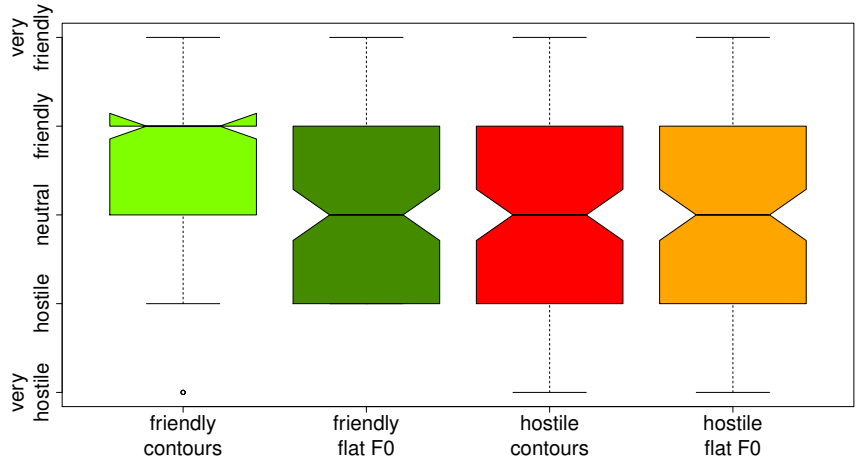

figure 9. Deuxième évaluation : résultats d'étude perceptive sur les audio aux contours prosodique contrôlés précisément. En vert clair, le contour amical, en vert foncé, juste la F_0 amicale, en rouge, le contour hostile et en orange, la F_0 hostile

et un audio avec un contour prosodique hostile ((*F*$_0$ moyenne et contour prosodique calculé sur Spike) dont les résultats sont visibles dans la figure 9.

De nos résultats, un test de Shapiro-Wilk nous a indiqué une distribution non normale des réponses donc nous avons fait un test de Man-Withney'U pour comparer la médiane des perceptions de chaque cas. En ce qui concerne la perception des contours prosodiques amicaux, nous avons observé un effet significatif par rapport aux contours plats hostiles ($p = 3.10^{-5}$) et aux contours prosodiques hostiles (p=0.003) ainsi qu'une forte présomption d'effet par rapport aux contours plats amicaux (p=0.011). Nous avons également mené une analyse plus détaillée pour chaque sujet sur l'évolution de sa perception entre un contour plat et un contour prosodique. Nous avons observé que les contours amicaux ont le meilleur consensus sur la perception de l'amicalité avec une faible variation dans les jugements. Cela n'est pas le cas pour les contours hostiles où la différence entre la présence et l'absence de contour ont été évaluées avec plus de différences dans les perceptions. Ces résultats restent en accord avec la littérature comme quoi la variation de *F*$_0$ peut influencer la perception de l'amicalité mais a un effet moindre sur l'hostilité.

5. Conclusion

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 multimododale 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.

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*.

Boersma P., Weenink D. (2017, March). *Praat: doing phonetics by computer [computer program]. version 6.0.27.* Consulté sur http://www.praat.org/

Cafaro A., Vilhjálmsson 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 *International conference on intelligent virtual agents.*

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 *Privacy, security, risk and trust (passat), 2012 international conference on and 2012 international conference on social computing (socialcom).*

Chollet M., Ochs M., Pelachaud C. (2014). From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In *International conference on intelligent virtual agents.*

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 *Proc. of lrec int. workshop on emotion.*

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 *2014 ieee international conference on acoustics, speech and signal processing (icassp).*

Dermouche S., Pelachaud C. (2016). Sequence-based multimodal behavior modeling for social agents. In *18th acm international conference on multimodal interaction.* ACM.

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 *Asian conference on machine learning.*

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 *International conference on intelligent virtual agents.*

Keltner D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of personality and social psychology*, vol. 68, n° 3.

Laskowski K., Edlund J., Heldner M. (2008). Learning prosodic sequences using the fundamental frequency variation spectrum. In *Proc. 4th international conference on speech prosody.*

Lee J., Marsella S. (2012). Modeling speaker behavior: A comparison of two approaches. In *Iva.*

Martínez H. P., Yannakakis G. N. (2011). Mining multimodal sequential patterns: a case study on affect detection. In *Proceedings of the 13th international conference on multimodal interfaces.*

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. *IEEE Transactions on Affective Computing*, vol. 3, n° 1.

Nicolle J., Bailly K., Chetouani M. (2015). Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. , vol. 6.

Ochs M., Pelachaud C. (2012). Model of the perception of smiling virtual character. In *11th international conference on autonomous agents and multiagent systems-volume 1.*

Pentland A. (2004). Social dynamics: Signals and behavior. In *3rd international conference on developmental learning.*

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 *International workshop on intelligent virtual agents.*

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 *Ieee international conference on computer vision workshops.*

Savran A., Cao H., Nenkova A., Verma R. (2014). Temporal bayesian fusion for affect sensing: Combining video, audio, and lexical modalities. *IEEE transactions on cybernetics*, vol. 45, n° 9.

Scherer K. R. (2005). What are emotions? and how can they be measured? *Social science information*.

Srikant R., Agrawal R. (1996). Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology—EDBT'96*.

Truong K., Heylen D., Chetouani M., Mutlu B., Salah A. A. (2015). Erm4ct '15: Proceedings of the international workshop on emotion representations and modelling for companion technologies.

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*, vol. 27, n° 12.

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. *IEEE Transactions on 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 *International conference on intelligent virtual agents.*