Analyse de signaux sociaux multimodaux : application à la synthèse d'attitudes sociales chez un agent conversationnel animé

Thomas Janssoone

EDITE - ED 130

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)
présentée et soutenue publiquement par

Thomas Janssoone

le 13 février 2018

Analyse de signaux sociaux multimodaux :

application à la synthèse d'attitudes sociales chez un agent conversationnel animé

Directeur de thèse : **Gaël Richard**
Co-encadrement de la thèse : **Chloé Clavel**
Co-encadrement de la thèse : **Kévin Bailly**

devant le jury composé de :

M. Pierre DE LOOR Professeur, LabSTICC, ENIB	Rapporteur
M. Alexandre PAUCHET Maître de conférences, LITIS, INSA Rouen	Rapporteur
Mme Catherine PÉLACHAUD Professeur, ISIR, Sorbonne Université	Examinateur
Mme Justine CASSELL Professeur, ArticuLab, Université Carnegie Mellon	Examinateur
Mme Magalie OCHS, Maître de conférences,LSIS, Aix-Marseille Université	Examinateur

Résumé

LORS d'une interaction, le comportement non-verbal apporte des informations sur l'état affectif de l'intervenant comme son attitude ou sa personnalité par exemple. Cela se traduit par des modulations dans l'utilisation de ses signaux sociaux : les variations dans les mouvements de tête, les expressions faciales ou la prosodie traduisent ces différents phénomènes affectifs. Désormais, l'utilisation d'agents conversationnels animés permet aux machines d'utiliser le même type de signaux sociaux. Ces agents peuvent ainsi améliorer la qualité de vie dans nos sociétés modernes s'ils proposent une interaction naturelle avec des utilisateurs humains. Pour cela, l'agent virtuel doit être capable d'exprimer différentes attitudes selon l'utilisateur, comme de la dominance pour un tuteur ou de la bienveillance pour un compagnon.

La littérature en sociologie et psychologie souligne que la dynamique dans l'usage des signaux sociaux contient une information importante pour l'expression de différents états affectifs. Les travaux présentés dans cette thèse proposent donc des modèles centrés sur la temporalité, élaborés à partir de signaux sociaux extraits automatiquement de corpus d'études, afin d'exprimer un phénomène affectif voulu. L'analyse de cette information est toujours effectuée dans un but de synthèse de comportements pour pouvoir l'utiliser lors de la génération d'agents conversationnels animés. Ainsi, une revue des bases de données existantes justifie l'élaboration, dans cette thèse, d'un corpus de travail composé d'allocutions présidentielles. Les vidéos de bonne qualité le composant permettent alors l'utilisation d'algorithmes pour évaluer automatiquement les signaux sociaux. Après un traitement des signaux sociaux extraits, des vidéos sont générées où un agent clone les allocutions. Cela permet d'évaluer et de comparer la perception d'attitude avec l'humain et avec l'agent virtuel comme protagoniste. Le modèle *SMART* utilise la fouille de données pour trouver des règles d'associations temporelles dans des corpus d'interactions. Il permet de trouver une information temporelle précise dans l'utilisation de signaux sociaux et de la lier avec une attitude sociale. La structure de ses règles permet également de transposer cette information pour synthétiser le comportement d'un agent virtuel. Des études perceptives viennent valider cette approche. Une collaboration internationale a abouti au modèle *SSN* qui se base sur de l'apprentissage profond et de la séparation de domaine. Il permet un apprentissage multi-tâche de plusieurs phénomènes affectifs simultanément et propose ainsi une méthode d'analyse de la dynamique des signaux employés.

Ces différentes contributions confirment l'intérêt de prendre en compte la temporalité dans la synthèse d'agents virtuels pour exprimer correctement certains phénomènes affectifs. Les perspectives proposent des pistes pour l'intégration de cette information dans des solutions multimodales.

Mots-clefs : <dynamique des signaux sociaux, fouille de données, synthèse d'agents virtuels>.

Abstract

DURING an interaction, non-verbal behavior reflects the emotional state of the speaker, such as attitude or personality. Modulations in social signals tell about someone's affective state like variations in head movements, facial expressions or prosody. Nowadays, machines can use embodied conversational agents to express the same kind of social cues. Thus, these agents can improve the quality of life in our modern societies if they provide natural interactions with users. Indeed, the virtual agent must express different attitudes according to its purpose, such as dominance for a tutor or kindness for a companion.

Literature in sociology and psychology underlines the importance of the dynamic of social signals for the expression of different affective states. Thus, this thesis proposes models focused on temporality to express a desired affective phenomenon. They are designed to handle social signals that are automatically extracted from a corpus. The purpose of this analysis is the generation of embodied conversational agents expressing a specific stance. A survey of existing databases lead to the design of a corpus composed of presidential addresses. The high definition videos allow algorithms to automatically evaluate the social signals. After a corrective process of the extracted social signals, an agent clones the human's behavior during the addresses. This provides an evaluation of the perception of attitudes with a human or a virtual agent as a protagonist. The *SMART* model use sequence mining to find temporal association rules in interaction data. It finds accurate temporal information in the use of social signals and links it with a social attitude. The structure of these rules allows an easy transposition of this information to synthesize the behavior of a virtual agent. Perceptual studies validate this approach. A second model, *SSN*, designed during an international collaboration, is based on deep learning and domain separation. It allows multi-task learning of several affective phenomena and proposes a method to analyse the dynamics of the signals used.

These different contributions underline the importance of temporality for the synthesis of virtual agents to improve the expression of certain affective phenomena. Perspectives give recommendation to integrate this information into multimodal solutions.

Keywords : <dynamic of social signals, sequence mining, virtual agent synthesis>.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction générale

Sommaire

LORS d'une interaction, les humains vont exprimer leur état émotionnel non seulement par le contenu verbal qu'ils emploient mais également par les signaux sociaux qu'ils utilisent. Ils vont ainsi moduler leur voix, varier leurs mouvements et les expressions de leur visage tout au long de l'interaction. L'étude de l'utilisation de ces signaux correspond au domaine du traitement du signal social qui cherche, dans le même temps, à comprendre et à modéliser les interactions sociales entre humains et à donner aux machines des capacités d'interactions similaires (Vinciarelli et al., 2012).

Ces progrès en traitement du signal social conduisent à l'utilisation d'agents conversationnels animés comme interfaces avec un ou plusieurs utilisateurs. La machine obtient ainsi la capacité d'exprimer des émotions, des attitudes ou d'autres états affectifs. Ces agents conversationnels animés sont des personnages virtuels qui peuvent dialoguer avec l'utilisateur et exécuter un ou plusieurs programmes si nécessaire. 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'agent la plus fluide et la plus naturelle possible. Pour cela, le contrôle des "signaux sociaux" émis par le personnage virtuel, comme ses expressions faciales ou sa synthèse vocale, permet de lui faire exprimer différentes at-

titudes envers l'utilisateur, comme de la dominance pour un tuteur ou de la bienveillance pour un compagnon.

Ce chapitre va permettre d'expliquer le contexte de cette thèse afin de bien définir son domaine d'application. Cela permet également de justifier sa problématique et les questions de recherche étudiées dans ces travaux.

1.1 Le domaine du traitement du signal social

Au-delà de l'intelligence abstraite, mesurée par le quotient intellectuel, les humains font preuve d'une intelligence sociale lors de leurs interactions avec le reste du monde (Albrecht, 2006). Elle traduit la capacité de reconnaître et d'exprimer différents signaux sociaux et comportements afin de bien s'entendre avec les autres tout en arrivant à collaborer avec eux (Vinciarelli et al., 2009b). Ces signaux sociaux sont ainsi la manifestation d'un état socio-émotionnel correspondant à la situation et à l'interaction en cours. Divers indices non verbaux permettent de les exprimer dont les expressions faciales, les mouvements du corps, les gestes ou la prosodie, illustrés sur la figure 1.1, (Knapp et al., 2013). Ces signaux vont servir à développer la conscience sociale, l'empathie envers les autres, sur laquelle va ensuite s'établir un lien social qui va assurer l'efficacité d'une interaction (Goleman, 2006).

Le comportement non-verbal joue donc un rôle essentiel dans la communication sociale car les signaux utilisés vont indiquer différentes émotions, attitudes, et tous les autres phénomènes affectifs qui seront détaillés dans la partie 1.2. Juste en observant les expressions faciales d'une personne, il est possible de savoir si elle est plutôt joyeuse, en colère ou triste. De même, le ton de la voix lors d'un échange entre deux personnes va indiquer leur intérêt pour la conversation, leur engagement. La perception sociale d'une interaction dépend donc tout autant du contenu non verbal que du message échangé (Pentland,

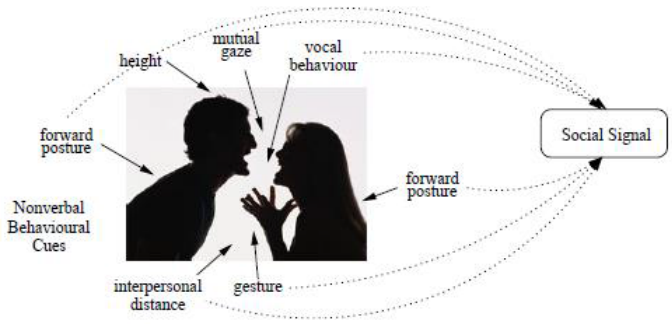

FIGURE 1.1 – Exemple de signaux sociaux et indices non verbaux, image issue de Vinciarelli et al. (2009b)

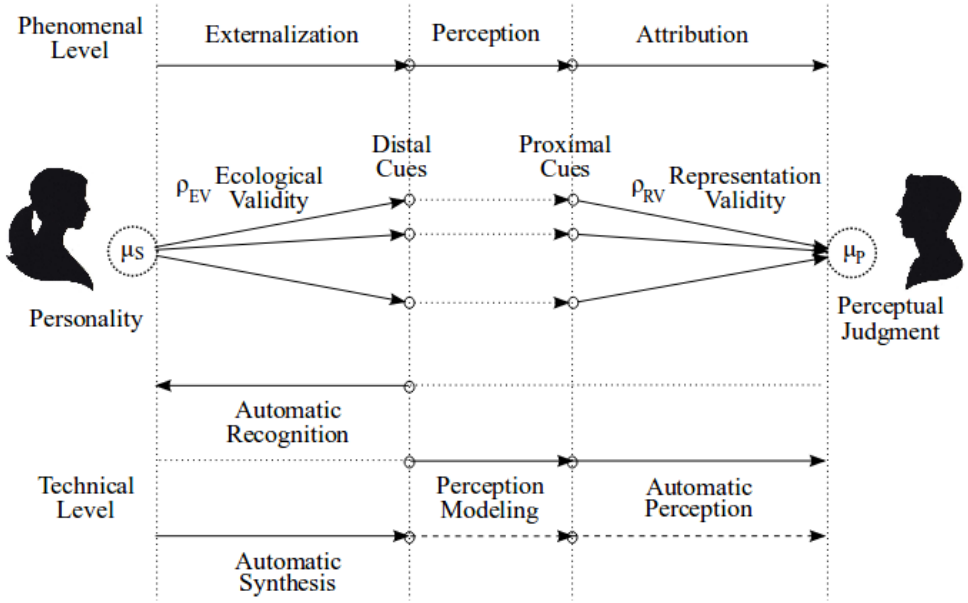

FIGURE 1.2 – Illustration du modèle de la lentille de Brunswik (1956) issu de Vinciarelli and Mohammadi (2014)

2004). La capacité à décoder correctement ces signaux, c'est-à-dire le niveau d'intelligence sociale, est donc essentiel à la réussite de toute interaction, ce qui vaut aussi pour une machine, initialement socialement inapte.

Ce processus d'échange implicite peut être modélisé par la lentille de Brunswik (1956), illustré dans la figure 1.2 : des indices non-verbaux sont exprimés par un intervenant, l'émetteur, et décodés par un autre intervenant, le récepteur. Ce dernier va en conclure des jugements subjectifs sur l'état de son interlocuteur. Les indices non-verbaux comportent plusieurs signaux dont le regard, les expressions faciales (mouvements des sourcils, pommettes, lèvres...), les mouvements de tête, du corps, la prosodie (ton de la voix, débit, volume...). Ce modèle cognitif décrit l'externalisation et l'attribution de caractéristiques socialement pertinentes pour décrire des interactions humain-humain ou humain-machine plus récemment. Lors de l'externalisation, l'émetteur, que ce soit un humain ou une machine, va moduler des signaux sociaux ce qui va refléter son état socio-émotionnel. Le récepteur va, pour l'attribution, capter une partie de ces indices et les analyser pour en déduire un jugement de l'émetteur.

Ce modèle permet de différencier deux champs d'applications du domaine du signal social (Vinciarelli et al., 2009b).
Le premier concerne la *reconnaissance* à partir d'indices non-verbaux qu'un ordinateur peut quantifier. Il s'agit de déduire l'état socio-émotionnel à partir des différents signaux

sociaux employés par le sujet étudié. Un ou plusieurs algorithmes vont ainsi modéliser ces signaux pour inférer des probabilités sur l'état du protagoniste. Cela peut être son émotion, son niveau de stress ou son implication dans l'échange (Clavel et al., 2008; Aigrain et al., 2016b).
La seconde application est liée à la *perception* de ces signaux par un observateur extérieur qui va lui permettre d'établir un jugement. Ces indices doivent donc être quantifiables par un humain : sur la voix par exemple, l'oreille n'entend pas la fréquence fondamentale (indice acoustique) mais est sensible à la hauteur (indice perceptif) qui correspond aux aigus/graves.
Ces deux domaines vont se rejoindre lors de la synthèse de comportements d'un agent car il faut que les signaux reconnus lors de l'élaboration du modèle soient bien perçus par un observateur extérieur comme exprimant le phénomène affectif désiré.

Ainsi, l'étude des interactions humaines a permis de donner à des machines la capacité de reconnaître et d'exprimer différents phénomènes affectifs. Cela a été rendu possible par l'élaboration de nombreux corpus d'études. Il s'agit généralement de bases de données audio-vidéos contenant des interactions entre deux ou plusieurs intervenants. Ces données sont généralement annotées par plusieurs observateurs extérieurs qui indiquent ainsi leurs perceptions de différents états socio-émotionnels exprimés. Différentes caractéristiques peuvent être extraites des fichiers audio-vidéo comme cela est illustré dans la section 3.2.

Un modèle informatique d'apprentissage peut être élaboré à partir de ces données en utilisant les annotations comme *classe*. Le modèle va chercher des associations entre des annotations et les différentes caractéristiques présentes en fusionnant ces dernières. Les annotations ou jugements humains peuvent être utilisés directement ou un travail peut être effectué afin d'assurer leur pertinence pour chaque classe. Par exemple, Aigrain et al. (2016a) proposent ainsi le calcul d'un poids de consensus inter-annotateur lors d'une étude sur l'évaluation du stress. Ils montrent ainsi de meilleurs scores de reconnaissance avec différents algorithmes d'apprentissage en intégrant cette information lors de l'entraînement des différents modèles. Ce travail souligne la sensibilité des modèles aux jugements humains sur lesquels ils se basent.
Dans le cadre d'une tâche de *reconnaissance*, le modèle va alors proposer, selon les données en entrée, un score de probabilité, une évaluation, de la présence de l'état socio-émotionnel étudié.
Pour une tâche de *génération*, le modèle va, selon une entrée donnée, prédire pour chaque modalité les modulations à apporter pour exprimer un état socio-émotionnel voulu, illustré dans la figure 1.3.

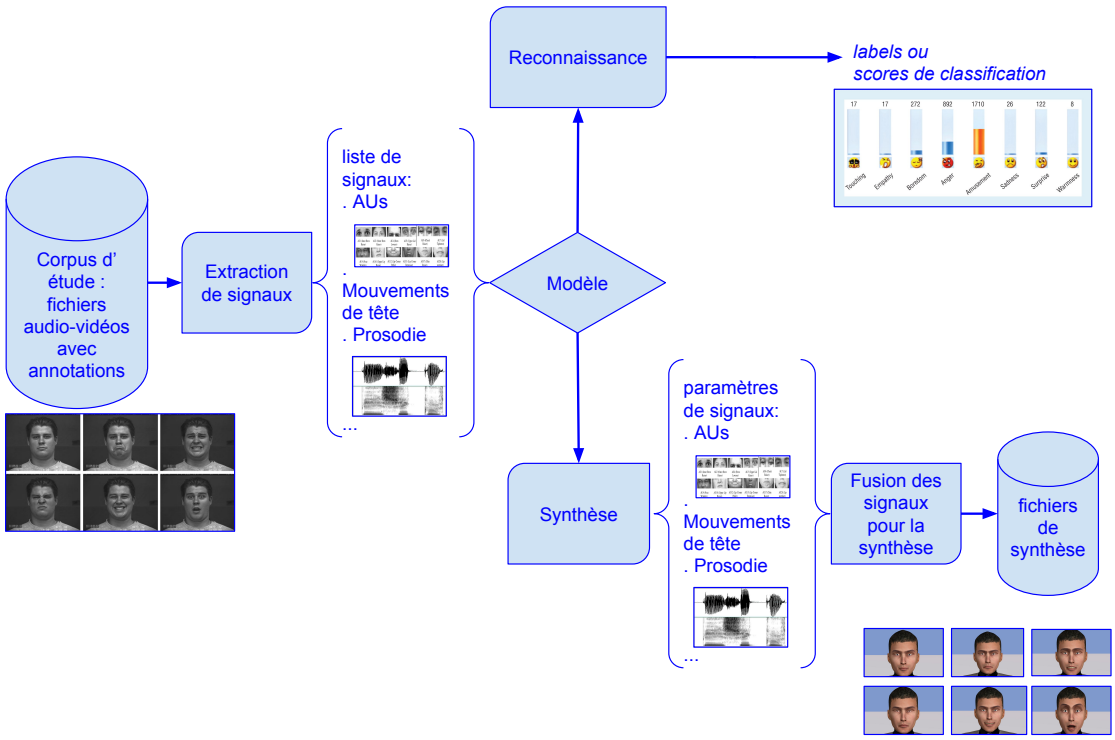

FIGURE 1.3 – Principe général de la création d'un modèle à partir d'un corpus de travail : en haut la tâche de *reconnaissance*, en bas la *génération* de comportements d'agents. Les classes d'apprentissage sont illustrées ici par les différentes émotions. Elles sont jouées par l'acteur à gauche, les probabilités d'appartenance à chacune apparaissent dans les jauges en haut, et l'agent les exprime en bas.

1.2 Les phénomènes affectifs

La littérature en psychologie définit différents phénomènes affectifs qui sont tout autant de champs d'applications au domaine du traitement du signal social. Ces définitions ont fait l'objet de nombreux débats et évolutions durant les dernières décennies. En suivant le consensus général, les spécifications proposées par Scherer (2005) sont d'abord détaillées puis d'autres études permettent d'ajuster le cadre d'application de cette thèse.

Dans ses recherches sur les émotions, Scherer (2005) propose un ensemble de critères afin de distinguer les phénomènes affectifs suivants : les *émotions*, les *préférences*, les *attitudes*, les *humeurs*, les *dispositions affectives* et les *postures interpersonnelles*.

- Focalisation sur l'événement : y a-t-il un événement déclencheur du phénomène affectif ? Le phénomène peut être la réponse à un événement interne (par exemple un souvenir) ou externe (par exemple un orage) ou être l'expression d'une décision

stratégique ou intentionnelle. L'événement est alors le déclencheur du processus d'évaluation (ou traitement d'évaluation, en anglais *appraisal*).

- Produit de l'évaluation : le phénomène exprime-t-il la pertinence d'un événement ? Certains phénomènes émotionnels expriment également l'intérêt suscité par l'événement déclencheur. Klaus Scherer différencie également l'*évaluation intrinsèque* qui est indépendante des besoins courants de l'*évaluation transactionnelle* qui reflète les buts et les désirs.
- Synchronisation de la réponse : quel est le degré d'engagement de l'organisme à la réponse à l'événement ? Le processus d'évaluation peut s'exprimer par une réponse synchronisée exprimée par l'organisme.
- Rapidité de changement : le phénomène varie-t-il rapidement ? Des événements et leurs évaluations changent très rapidement et peuvent modifier le phénomène affectif pour le réévaluer et l'adapter suite à ces nouvelles informations.
- Impact comportemental : le phénomène modifie-t-il beaucoup le comportement en cours ? Certains phénomènes affectifs ont un fort impact sur le comportement qui suit leur expression. Ils peuvent modifier le comportement en cours, voir l'interrompre complètement, créant alors de nouveaux buts et objectifs. Ces phénomènes peuvent par exemple modifier les expressions vocales et faciales avec un impact sur la communication et la relation sociale en cours
- Intensité : quelle est l'importance du phénomène dans l'adaptation comportementale, la réponse proposée ? L'intensité de l'expression de la réponse va ainsi différencier certains phénomènes affectifs entre eux.
- Durée : quelle est la durée de la manifestation du phénomène ? Certains phénomènes, par la mobilisation importante de leur expression, vont avoir alors une durée plus courte afin de préserver les ressources de l'organisme.

Chaque critère permet de caractériser les phénomènes affectifs et de souligner les différences entre eux. Klaus Scherer les utilise donc pour proposer les définitions suivantes, résumées dans le tableau 1.1.

Les *préférences* sont ainsi des jugements relativement stables concernant l'appréciation d'un stimulus ou le fait de le préférer par rapport à d'autres objets ou stimuli.

Les *attitudes* sont des croyances relativement stables et des prédispositions envers des objets ou personnes spécifiques. Elles n'ont pas besoin d'être déclenchées par les estimations de l'événement même si elles peuvent être modulées par celui ci.

Les *humeurs* sont considérés comme des états affectifs diffus, caractérisés par la prédominance durable de certains sentiments, affectant alors l'expérience et le comportement d'une personne. Elles sont souvent indépendantes de tout événement, de faible intensité mais d'une durée très importante (heures voir jours).

phénomène affectif	focalisation sur l'événement	évaluation intrinsèque	évaluation transactionnelle	synchronisation de la réponse	rapidité de changement	impact comportemental	intensité	durée	exemples
préférences	- -	+ +	.	- -	- -	.	-	.	aimer, détester
attitudes	- -	-	-	- -	-	-	.	+	détester, évaluer, désirer
humeurs	-	.	-	-	.	+	.	+	gaie, sombre, dépressive
dispositions affectives	- -	-	- -	- -	- -	-	-	+ +	nerveux, inquiet, hostile
postures interpersonnelles	+	-	-	-	+ +	+	.	.	poli, froid, chaleureux
émotions esthétiques	+	+ +	-	+ +	+	-	-/.	-	être remué, intimidé, béatitude
émotions utilitaires	+ +	.	+ +	+ +	+ +	+ +	+	-	colère, peur, joie

TABLEAU 1.1 – Tableau récapitulatif des différents phénomènes affectifs définis par Scherer (2005) (inspiré de Piolat and Bannour (2008)).
L'importance de chaque caractéristique est symbolisée ainsi :
- - très faible, - faible , . moyenne, + grand, ++ majeur.

Les *dispositions affectives* sont des traits de la personnalité et des tendances à l'action (nervosité, anxiété, irritabilité, ...). Cela décrit la propension d'une personne de ressentir certaines humeurs plus fréquemment ou sa façon de réagir (ou sur-réagir) à certains type d'émotions.

Les *postures interpersonnelles* sont 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 poli, distant, froid, chaleureux, compassionnel, dédaigneux).

Les *émotions esthétiques* impliquent une adaptation immédiate à un événement sans évaluation de son impact par rapport au but. Dénuées de considérations fonctionnelles, elles sont produites par l'appréciation des qualités intrinsèques d'une œuvre d'art ou d'un phénomène naturel par exemple.

Les *émotions utilitaires* facilitent l'adaptation à un événement et vont avoir un impact fort sur le bien être de la personne. Cela regroupe les émotions dites "classiques" (peur, joie, dégoût, surprise, tristesse ...). En préparant à l'action (fuite, affrontement) ou à la motivation (joie, fierté), les émotions utilitaires s'avèrent être des réactions urgentes et d'intensités importantes, demandant alors une mobilisation rapide et conséquente de l'organisme.

Il est intéressant de relier ces définitions aux recherches d'Argyle (1975) qui liste cinq types de fonctions à la communication non-verbale. Elle peut servir à soutenir des *rituels sociaux* comme les poignées de mains ou accompagner l'apparence dans la *présentation de soi* pour souligner son appartenance à un groupe social par exemple. Le comportement non-verbal va avoir aussi un rôle d'*accompagnement et de support dans le discours* en régulant la conversation via des vocalisations non-verbales ou des hochements de tête par exemple. L'*expression d'émotions* va également se faire par le langage non-verbal : grimace de dégoût ou sursaut de peur. Argyle souligne qu'en dehors des réactions physiologiques

spontanées, certaines émotions se sont développées au fil de l'évolution comme un signal social, toujours spontané, mais avec une utilité communicative. Ainsi, la peur ou la colère peuvent servir à exprimer la présence d'un danger. L'expression d'émotions peut être également utilisée délibérément pour illustrer le discours sans que celle-ci ne soit réellement ressentie. Enfin, le comportement non-verbal va également permettre de *communiquer des attitudes interpersonnelles* : l'établissement et le maintien d'une relation avec autrui telle que l'amicalité.

C'est ce dernier aspect qui a été retenu dans le cadre de cette thèse, plus précisément *l'établissement d'une relation avec un utilisateur*. Pour les agents conversationnels animés, il s'agit d'un point essentiel pour améliorer l'interaction avec les humains. En effet, selon le but recherché, l'agent doit pouvoir établir différentes relations avec l'utilisateur : un tuteur virtuel pour de l'aide aux devoirs n'aura pas la même compassion ou chaleur qu'un compagnon quotidien chargé d'aider à la vie de tous les jours (Truong et al., 2015). Ces notions, par rapport aux définitions de Scherer (2005), se trouvent alors au niveau des postures interpersonnelles : il s'agit de faire des agents conversationnels animés chaleureux ou froid, dominant ou soumis.

Dans le but de pouvoir construire les modèles présentés dans cette thèse, des classes d'apprentissages doivent être définies comme cela a été illustré dans la figure 1.3. Pour les déterminer clairement, des études issues du domaine de la psychologie et de la sociologie vont maintenant être passées en revue. Elles vont permettre de détailler ces classes d'apprentissages, nécessaire pour la machine, mais aussi d'obtenir de l'information a priori sur les signaux sociaux jouant un rôle dans l'expression d'attitudes.

1.3 Application : les attitudes interpersonnelles

Il existe de nombreuses définitions dans la littérature où une attitude (*stance* en anglais, qui peut également se traduire par *posture*) est vue comme un jugement, un positionnement (Chindamo et al., 2012; Burgoon et al., 1984). En particulier, Du Bois (2007) propose ainsi un modèle en triangle suivant la définition : "*Stance is a public act by a social actor, achieved dialogically through overt communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects, with respect to any salient dimension of the sociocultural field*". Cela souligne le caractère dynamique de l'expression des attitudes par cette notion d'acte consistant à se positionner tout au long de l'interaction.
Cette communication des attitudes interpersonnelles comme l'établissement d'une relation peut se retrouver à l'intersection des définitions de Scherer (2005) entre les *attitudes* et les *postures interpersonnelles*. Les critères définis par Scherer (voir tableau 1.1) soulignent

alors son impact qui va s'exprimer via le comportement non-verbal pendant un temps conséquent.
Ces définitions soulignent donc que, par rapport aux autres états socio-émotionnels comme la personnalité, les émotions ou l'humeur, les attitudes interpersonnelles vont se révéler au fil de l'interaction par la modification du comportement d'un individu. Elles doivent donc se retrouver dans les modifications du comportement non verbal de l'intervenant via des variations dans l'utilisation de plusieurs signaux sociaux au fil du temps.

En effet, la dynamique des signaux sociaux employés va apporter des informations nécessaires à la compréhension de cette attitude. Keltner (1995) en illustre l'importance avec l'exemple suivant : un long sourire va être révélateur d'amusement tandis qu'un regard fuyant suivi d'un sourire contraint sera signe d'embarras. De même, dans ses recommandations pour l'étude des attitudes dans la communication, Chindamo et al. (2012) insiste sur le fait que les attitudes sociales se construisent au fur et à mesure des tours de paroles. L'enchaînement de ces signaux dans le temps apporte donc des informations sur ces attitudes et leur construction. Des modèles prenant en compte la dynamique des signaux sociaux sont donc particulièrement adaptés a l'étude de cet état socio-émotionnel.

Les modèles élaborés dans cette thèse ont été axés sur l'étude de la dynamique de signaux sociaux pour analyser et évaluer différents phénomènes affectifs. Ils sont certainement généralisables à d'autres problématiques mais les travaux présentés ici ont été appliqués à *trouver le lien entre l'expression d'attitudes sociales et la dynamique des signaux utilisés afin de trouver de l'information pertinente pour permettre la synthèse de cette même attitude chez un agent virtuel*.
Les travaux présentés ici vont par ailleurs se limiter à l'intra-synchronie d'un seul des protagonistes de l'utilisation de signaux sociaux multimodaux. Cela signifie qu'un seul des individus est étudié là où l'inter-synchronie correspondrait à l'étude des signaux de l'ensemble des intervenants. En effet, valider ce modèle "individuel" est une étape nécessaire avant de l'étendre à un modèle "collectif", plus complexe à analyser et interpréter.

Comme cela est illustré dans la figure 1.3, les modèles informatiques ont besoin d'une mesure des attitudes sociales afin d'établir les classes d'apprentissages. Une revue de la littérature en psychologie et sociologie va permettre d'établir cette mesure mais également d'avoir des connaissances à priori afin de valider les approches proposées ou d'utiliser des signaux sociaux pertinents pour réduire les temps de calcul des modèles.

1.3.1 Représentation et mesures d'attitudes sociales

Tout comme il existe de nombreuses nuances dans les définitions des attitudes interpersonnelles, plusieurs représentations permettent de les mesurer. Ces mesures ont ini-

tialement été mises en place pour améliorer la compréhension, le diagnostic et le traitement de diverses pathologies psychiatriques. Elles ont été largement débattues pendant la deuxième moitié du XX^{ème} siècle mais le modèle du circomplexe interpersonnel est aujourd'hui le plus répandu. Il a été défini par Leary (1958) et a longtemps fait référence dans la mesure des attitudes.

Il s'agit de proposer une représentation selon deux dimensions d'un espace interpersonnel donné (besoins, valeurs, traits, problèmes, ...) en l'organisant selon un cercle modélisant l'étendue de ses nuances (Wiggins, 1979; Kiesler, 1996; Horowitz et al., 2006). Il est ainsi possible de mesurer selon ces coordonnées les variables étudiées en les ordonnant dans cet espace continu. Chaque axe mesure ainsi une modalité fondamentale des interactions humaines : l'un représentant le pouvoir, (*agency* en anglais), lié plutôt à la différenciation de l'individu, le second étant la communion, (*communion* en anglais), lié à la connexion aux autres. Le pouvoir interpersonnel relate les idées de dominance, pouvoir, hiérarchie et contrôle. La communion reflète l'amour, l'affiliation, l'union et l'amicalité. Une projection proche des extrémités des axes indique un message clair et fort là où une position centrale indique plus de neutralité.

Argyle (1975) propose ainsi une représentation bi-dimensionelle des attitudes, illustrée dans la figure 1.4 qui sera retenue dans le reste de cette thèse. La dimension liée à la *communion* figure ici l'*amicalité*, une valeur positive exprimant de l'amicalité, une valeur négative correspond à de l'hostilité. Celle du *pouvoir* est liée à la *dominance*, une valeur positive indiquant une supériorité sociale, une valeur négative représentant une infériorité sociale.

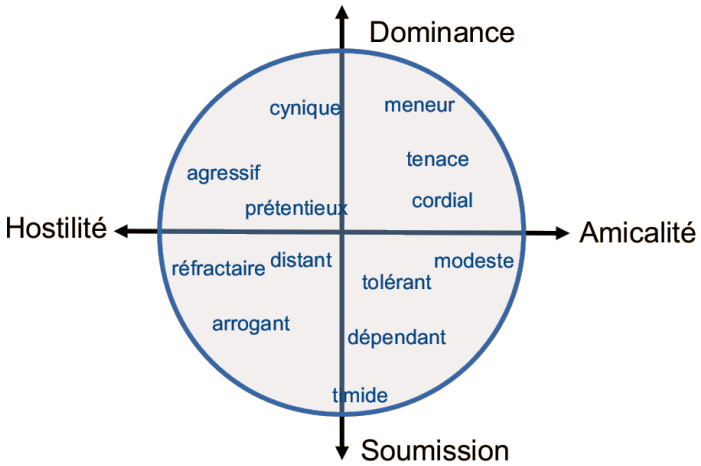

FIGURE 1.4 – Représentation du circomplexe interpersonnel, défini par Argyle

Cette représentation en deux dimensions suppose une distribution uniforme et continue des variations d'attitudes selon la position sur le cercle les représentant. Wiggins (1979) y fait ainsi référence comme une taxonomie du domaine interpersonnel étudié. L'ensemble des attitudes possibles est "encodé" selon le positionnement dans ce référentiel. Néanmoins, il reste possible de découper cette représentation en un nombre de sous catégories variables selon l'étude voulue. La figure 1.5 montre ainsi la découpe des attitudes sociales selon 4, 8, 16 segments. Ces différents segments ou secteurs ont permis d'aboutir à un certain consensus dans les dimensions définies mais surtout ont permis d'établir des mesures assez précises des attitudes.

Ce modèle du circomplexe interpersonnel est particulièrement adapté à la quantification précise de phénomène socio-émotionnel dont les attitudes. Locke (2006) illustre dans sa revue l'utilisation du circomplexe pour des études en psychopathologie comme un instrument de mesure précis. Un certain nombre d'échelles de jugement sont utilisées pour permettre d'effectuer un placement sur le circomplexe (voir figure 1.6). Parmi celles ci, le

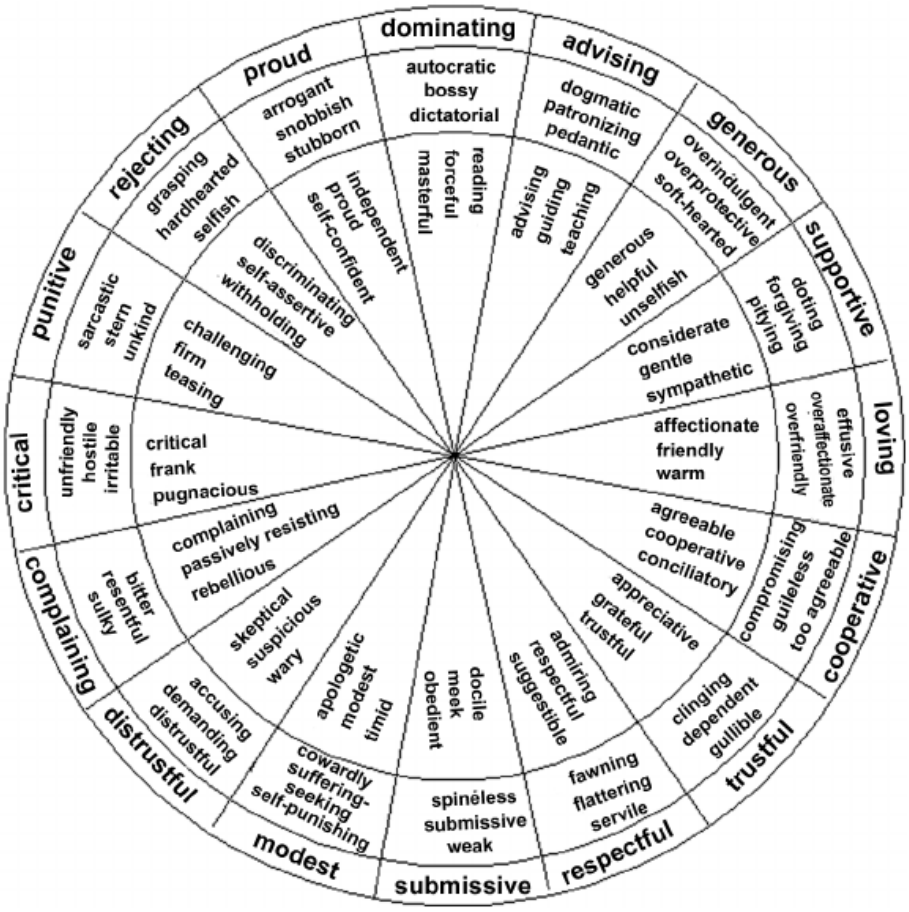

FIGURE 1.5 – Représentation du circomplexe interpersonnel, avec les différentes tailles de segments (4, 8 et 16), issue de Freedman et al. (1951)

Construct:	Interpersonal values	Impact messages	Interpersonal traits	Interpersonal problems	Social support transactions
Measure:	**CSIV**	**IMI-C**	**IAS-R**	**IIP-C**	**SAS-C**
Angle					
0°	Communal	Friendly	Warm-Agreeable	Overly Nurturant	Nurturant
45°	Agentic and Communal	Friendly-Dominant	Gregarious-Extraverted	Intrusive	Engaging
90°	Agentic	Dominant	Assured-Dominant	Domineering	Directive
135°	Agentic and Separate	Hostile-Dominant	Arrogant-Calculating	Vindictive	Arrogant
180°	Separate	Hostile	Cold-hearted	Cold	Critical
225°	Submissive and Separate	Hostile-Submissive	Aloof-Introverted	Socially Avoidant	Distancing
270°	Submissive	Submissive	Unassured-Submissive	Nonassertive	Avoidant
315°	Submissive and Communal	Friendly-Submissive	Unassuming-Ingenuous	Exploitable	Deferential

Legend and Source(s):
Circumplex Scales of Interpersonal Values (CSIV; Locke, 2000).
Octant Scale Impact Message Inventory (IMI-C; Schmidt, Wagner, & Kiesler, 1999).
Interpersonal Adjective Scales-Revised (IAS-R; Wiggins, 1995).
Inventory of Interpersonal Problems-Circumplex (IIP-C; Alden et al., 1990; Horowitz et al., 1988).
Support Actions Scale-Circumplex (SAS-C; Trobst, 2000).

FIGURE 1.6 – Cinq mesures basées sur le circomplexe interpersonnel, source Gurtman (2009)

Questionnaire Interpersonnel ("*Interpersonal Adjective Scales-Revised*") de Wiggins (1979) reste la mesure principale. Il a été construit pour suivre au mieux le circomplexe interpersonnel et fait office de référence. Elle consiste en huit échelles (ou octants) selon huit adjectifs (chaleureux, timide, ...). Ces différentes échelles permettent d'évaluer l'attitude d'une personne sans demander directement le jugement selon les deux axes. Cela permet de lisser le résultat et d'éviter un biais lié au sens des mots employés lors de l'appréciation.

Ainsi, à partir de ces mesures, Trapnell and Broughton (2006) proposent un questionnaire précis afin d'évaluer efficacement une attitude sociale composé de douze échelles de Likert-5 pour mener les évaluations. Il s'agit de douze adjectifs correspondant au circomplexe de Wiggins de représentation des attitudes : *Sûr de soit ("Assertive", AS : 90°), Dominant ("Dominant", DO : 130°), Manipulateur ("Manipulative",MA : 150°), Insensible ("Coldhearthed", CO : 180°), Réservé ("Aloof", AL : 210°), Introverti ("Introverted", IN : 240°), Timide ("Timid", TI : 270°), Déférent ("Deferent", DE : 300°), Agréable ("Agreeable", AG : 330°), Prévenant/Épanouissant ("Nurturant", NU : 0°), Chaleureux ("Warm", WA 30°) et Extraverti ("Extraverted", EX : 60°)*. Les formules 1.1 et 1.2 permettent de calculer les scores dans la représentation de Wiggins.

$$DOM = (DO + AS + EX) - (DE + TI + IN) \tag{1.1}$$

$$AMI = (WA + NU + AG) - (MA + CO + AL) \tag{1.2}$$

Dans la suite de cette thèse, les attitudes sociales ont donc été mesurées selon ces deux dimensions du circomplexe d'Argyle qui fait référence dans la communauté (Chollet et al., 2014; Ravenet et al., 2013; Pecune et al., 2014; Marsella et al., 2010; Zhao et al.,

2016). Que ce soit pour l'annotation de corpus (voir section 3.4) ou lors des évaluations perceptives des modèles (voir section 4.3), le questionnaire établi par Trapnell a été adapté en utilisant un facteur de normalisation pour garder ces scores entre -2 et 2 comme cela a déjà été fait dans les travaux de Pecune et al. (2016a).

1.3.2 Les signaux sociaux liés à la dominance et à l'appréciation

Afin de valider les modèles proposés dans cette thèse, une première approche consiste à les confronter aux résultats présents dans la littérature en sociologie et psychologie. En effet, les attitudes ont déjà été largement étudiées et les liens trouvés entre certains signaux sociaux et l'expression d'attitudes assureront la cohérence des résultats présentés.

Parmi les attitudes, l'aspect concernant l'expression de la dominance a été l'un des plus étudiés. Cela s'explique en partie par son impact dans la plupart des interactions sociales. Il peut ainsi faire allusion au rapport de force, au statut, au leadership tout comme à la dominance personnelle, au contrôle de soi. Dans cette thèse, c'est le premier aspect qui a été retenu car il est plus approprié aux attitudes sociales.

Dès ses premières recherches (1977), Henley (1995) affirme que la dominance est étroitement liée au comportement non-verbal et qu'ainsi, différents niveaux de dominance peuvent être exprimés par différents comportements non verbaux. Hall et al. (2005) proposent une analyse complète de la dimension verticale des relations sociales i.e. liée à la dominance, au pouvoir, au statut.

Tout d'abord, un fort lien entre attitudes et émotions est souligné. Ainsi, les individus dominants sont susceptibles d'utiliser peu d'expressions faciales liées à la peur et à la tristesse mais plus d'expression faciale de dégoût ou de colère. Hess and Thibault (2009); Knutson (1996); Tiedens (2000) montrent également que la joie, la colère et le dégoût sont liés à la dominance là où la soumission est connectée à la peur et la tristesse. Cela indiquerait que les individus dominants sont plus maîtres de l'expressivité de leur visage. En effet, le détail de plusieurs études fait par Hall ne montre pas de lien avéré entre la dominance et les sourires ou l'utilisation des sourcils. Le lien existe cependant avec l'intensité et l'expressivité du visage. Burgoon et al. (1984) ajoutent que les échanges de regards vont aussi asseoir la dominance (et l'amicalité).

Par ailleurs, un port de tête haut (*upward tilt*) est également révélateur de dominance (Carney et al., 2005). Mignault and Chaudhuri (2003) soulignent également que la perception de dominance est influencée par l'angle de la tête. Les individus dominants favoriseraient aussi un champs de vision dégagé là où un individu plus soumis en aura un plus restreint. Cela fait écho à l'association entre la dominance et une position plus élevée, un

maintien du corps droit (Robinson et al., 2008).

Dans le dialogue, les individus dominants utilisent moins de "um" ou de "ah", ont plus de succès dans les interruptions et ont un discours plus fluide et sûr (Carney et al., 2005; Burgoon and Le Poire, 1999). Les individus dominants auront également des tours de paroles plus longs (Burgoon et al., 1984). Ainsi, Wallbott and Scherer (1986) dans leur étude de la tristesse et de la colère montrent ainsi que la soumission est liée à un débit vocal plus lent, une hauteur et une intensité plus faibles et des gestes moins expansifs. La dominance correspond à une parole plus rapide, plus aiguë, et plus forte.

Wallbott and Scherer (1986) montrent aussi que les personnes plus dominantes utilisent des gestes plus amples. Carney et al. (2005) retrouvent ce résultat également dans les mouvements du corps, plus amples et plus ouverts chez les dominants, ce qui est confirmé par Burgoon and Le Poire (1999).

En ce qui concerne l'amicalité, Hess et al. (2000) indiquent que le sourire reste le marqueur principal de la bienveillance. Ils sont perçus comme chaleureux, bienveillants et amicaux (Hess et al., 2002). Deutsch et al. (1987) indiquent ainsi que les personnes souriantes laissent une impression plus joyeuse, détendue et plus polie là où l'absence de sourire est toujours perçue comme une intention négative (e.g. hostile). Cependant, il est important de noter qu'un sourire peut avoir différentes signification comme l'expression d'embarras, d'amusement ou de politesse (Ochs and Pelachaud, 2013; Ochs et al., 2013). Ainsi, parmi d'autres caractéristiques comme son amplitude ou sa symétrie, la durée et la vitesse d'un sourire permettent de repérer ces différents usages. Ravenet et al. (2013) montrent également que des expressions faciales liées à la colère ou au dégoût vont être perçues comme hostiles.

Les signaux qui sont liés à de la positivité (rire, sourire, hochement de tête), de l'intérêt (orientation du corps, échange de regards), de l'expressivité (voix plus aiguë et plus forte, variation de hauteur dans la voix, expressivité faciale) et une interaction plus fluide (tour de parole court, peu d'interruptions) vont également exprimer de l'amicalité (Burgoon and Le Poire, 1999).

Ces travaux sont résumés dans le tableau 1.2.

Signal	Influence la dominance	Influence l'appréciation
Expressions faciales liées aux émotions	plus d'expressions faciales de dégoût, joie ou colère, peu d'expressions liées à la peur et la tristesse	absence de sourire, expressions de dégoût et colère perçues comme hostile
Regard	échange de regard et champs de vision dégagé	échange de regards
Mouvements de tête	port de tête haut	hochement de tête
Dialogue	peu de "um" ou "ah" mais discours fluide et sûr	voix plus aiguë et forte, variations de hauteur, interaction fluide
Gestes	gestes amples et plus ouverts	-

TABLEAU 1.2 – Résumé des signaux sociaux liés à la dominance et à l'appréciation.

1.4 Problématique : la dynamique des signaux sociaux

1.4.1 Questions de recherche et objectifs

Cette thèse s'est déroulée dans le cadre du *Labex SMART* dont l'un des buts est d'encourager le transfert de connaissances entre différents laboratoires. L'*ISIR*, de l'université Pierre et Marie Curie, propose des solutions d'extractions automatiques d'expressions faciales sous forme d'unité d'actions du système *FACS* (*Facial action coding system*, plus détaillé dans la partie 3.2, référés ensuite par l'anglicisme *action units*). Dans le même temps, le *LTCI*, de Télécom ParisTech, offre une expertise en traitement de la parole. L'équipe *Greta* a permis d'avoir accès à une plate-forme complète pour la synthèse de comportements d'agents.
Ces connaissances ont été mises à profit pour extraire automatiquement les signaux sociaux utilisés dans cette thèse et apprendre à les calibrer pour la synthèse d'agents. Cet environnement a ainsi influencé les questions de recherches posées dans cette thèse grâce aux outils et expertises disponibles pour mener à bien ces travaux.

Cette thèse se focalise sur l'analyse de l'expression d'attitudes pour améliorer la synthèse de comportement pour un agent conversationnel animé. Elle se place alors dans un contexte où de nombreux travaux ont déjà été effectués.
Une littérature riche en sociologie et psychologie, présentée dans ce chapitre (voir section 1.3.2), liste des informations a priori sur les signaux liés à chaque attitude. Ces informations permettent de valider les résultats mais aussi de sélectionner des ensembles restreints de signaux sociaux pour limiter les calculs des modèles.
L'analyse des contributions précédentes montre que les modèles existants sur les attitudes peuvent s'intéresser au contexte de l'interaction comme l'étude de Pecune et al. (2016b)

sur la relation entre un agent et un utilisateur. D'autres modèles se concentrent sur l'analyse des caractéristiques de l'interaction avec un focus sur les signaux utilisés comme Chollet et al. (2013). Cette thèse s'intéresse plus à ce deuxième angle de recherche dont les contributions précédentes sont détaillées dans le chapitre 2. Enfin, des modèles récents (Youssef et al., 2015) intègrent ces deux composantes pour générer des agents adaptatifs.

Les travaux présentés ici ont pour but d'améliorer les connaissances sur ces caractéristiques grâce à des modèles informatiques. Cette thèse cherche à répondre à la question suivante :

Comment utiliser des signaux sociaux extraits automatiquement pour modéliser une dynamique propre à l'expression de comportements socio-émotionnels donnés ?

Cette question de recherche souligne deux aspects.
Le premier est qu'une information temporelle précise est souvent la partie manquante dans les modèles d'attitudes. Cependant, la littérature souligne que cette dynamique comporte une information importante dans l'expression de cet état émotionnel (voir section 1.3). Cette thèse va donc adapter différentes méthodes pour proposer des modèles d'expressions d'attitudes centrés sur la temporalité.
Le deuxième aspect concerne l'utilisation de signaux extraits grâce à des algorithmes ce qui permet une approche quasi-automatique. Jusqu'ici, la plupart des modèles pour la synthèse de comportements sont appris sur des données annotées manuellement par des observateurs humains. L'extraction des caractéristiques est alors fastidieuse et coûteuse. Le contexte de cette thèse a permis l'accès à des outils robustes d'extraction de signaux sociaux qui ont permis de minimiser l'intervention humaine dans la construction des modèles.

Les objectifs de ces travaux sont donc :

- Intégrer l'information temporelle dans les modèles de phénomènes affectifs.
- Adapter cette information aux contraintes de la synthèse d'agent à partir de signaux extraits automatiquement
- Proposer une méthodologie d'évaluation dans la conception de cette approche

Un certain nombre de restrictions ont été mises en place afin d'atteindre ces buts dans le temps alloué par cette thèse. Le cadre d'application se concentre sur l'analyse des signaux sociaux multi-modaux du visage (*action units*, mouvements de tête et prosodie) pour un seul intervenant d'une interaction dyadique.
Ces restrictions ont limité les corpus possibles pour correspondre à ce cadre d'étude tout en permettant l'extraction automatique des signaux sociaux, comme cela sera détaillé dans le chapitre 3.
Cela permet ensuite l'élaboration de deux modèles qui s'intéressent à l'étude d'un protagoniste dans l'interaction. Lors de la conclusion de chaque modèle, les ajustements à

apporter pour leurs utilisations avec d'autres signaux sociaux mais aussi pour prendre en compte l'information dyadique sont discutés.

1.5 Liste des contributions

LES travaux de cette thèse se sont donc placés dans le cadre de l'apprentissage par corpus pour la génération de comportement d'agents conversationnels animés. Ainsi, ils ont suivi le schéma de fonctionnement "classique" qui est illustré dans la figure 1.3 : des données en entrée aux fichiers de synthèse en passant par les modèles d'analyse. Les contributions de cette thèse reflètent ce processus et correspondent donc à chacune des étapes. Elles peuvent être divisées en deux grandes parties : données et modèles.
Les premières contributions sont donc sur les données :

- *Élaboration d'un corpus multimodal d'études annoté en attitude interpersonnelles et en prise de parole en public*
- *La mise en place d'un processus de traitement de données extraites automatiquement en vue d'une tâche de synthèse*

Elles sont détaillées dans le chapitre 2. Avec une analyse des corpus existants, les premières études mettent en avant un certain nombre de limitations. Elles sont en partie liées aux contraintes choisies dans le cadre de cette thèse : l'utilisation d'algorithmes pour extraire les signaux sociaux sans intervention humaine nécessite une certaine qualité dans les données à étudier ainsi qu'un certain cadrage des intervenants.
Un corpus dédié a été mis en place. Il fournit ainsi une base de travail contenant des vidéos de bonne qualité, permettant une extraction automatique des signaux fiable, avec des annotations en attitudes sociales. De plus, il sera enrichi au fil des années à venir. Il offre ainsi à la communauté des vidéos de très bonne qualité, annotées finement en attitudes sociales, pour des allocutions en public.
Ce corpus a mené à l'élaboration d'un processus automatique d'extraction et de traitement de signaux sociaux focalisé sur le but final : une tâche de synthèse sur un agent virtuel. Les recherches ont aussi porté sur la méthodologie de traitement des signaux extraits automatiquement pour cette tâche de synthèse. Passer directement du corpus annoté en comportement aux modèles de génération (sans passer par des annotations manuelles en signaux sociaux) diffère des traitements classiques pour des tâches de classification. De plus, l'extraction automatique comporte des risques d'erreurs qu'il a fallu maîtriser.

Deux angles d'analyse ont ensuite été explorés, permettant la construction de deux modèles qui constituent deux autres contributions :

- *Une méthode d'analyse et de synthèse basée sur de la fouille de données :* SMART *pour* Social Multimodal Association Rules with Timing *qui trouve des règles d'associations temporelles entre les signaux sociaux et les lie à l'expression d'attitudes sociales.*
- *Une méthode d'analyse et de synthèse basée sur de l'apprentissage profond :* SSN *pour* Social Separation Network *qui cherche dans la dynamique des signaux sociaux des représentations propres à une tâche et des représentations partagées par plusieurs.*

Ces solutions sont présentées dans les chapitres 4 et 5. Les deux modèles intègrent la dynamique temporelle des signaux sociaux et l'utilisent pour discriminer différents états affectifs. Les différentes solutions trouvent ainsi des représentations caractéristiques dans les données étudiées qui peuvent ensuite être synthétisées avec un agent virtuel.
Les résultats des deux méthodes sont cependant très différents. Par exemple, les représentations trouvées par la fouille de données sont plus facilement interprétables par un humain. Néanmoins, ils soulignent tous l'intérêt de prendre en compte cette dynamique et la faisabilité d'utiliser des signaux extraits automatiquement. La conclusion de ce manuscrit discute des limites actuelles de ces modèles et propose des pistes pour les dépasser afin de compléter la génération de comportement d'agents conversationnels avec l'expression d'un état affectif choisi.

1.5.1 Liste des publications lors de cette thèse

Cette thèse a donné lieu a plusieurs publications nationales et internationales, listées plus bas, ainsi qu'à des communications comme au GDR-ISIS [1]. Des échanges ont également eu lieu lors de l'école d'été *ISSAS 2016* [2] et lors d'une collaboration au sein de l'ICT [3].

- Janssoone, T. (2015). Temporal association rules for modelling multimodal social signals. In *proceedings of the International Conference on Multimodal Interaction (doctoral consortium)*
- Janssoone, T., Clavel, C., Bailly, K., and Richard, G. (2016a). Des signaux sociaux aux attitudes : de l´utilisation des règles d´association temporelle. In *proceedings of the WACAI 2016, Workshop . Affect . Compagnon Artificiel . Interaction*
- Janssoone, T., Clavel, C., Bailly, K., and Richard, G. (2016b). Using temporal association rules for the synthesis of embodied conversational agents with a specific stance. In *proceedings of the International Conference on Intelligent Virtual Agents*

1. http://www.gdr-isis.fr/index.php?page=reunion&idreunion=323
2. http://affective-sciences.org/home/education/summer-school-issas-2018/summer-school-issas/
3. http://ict.usc.edu/

- Janssoone, T., Clavel, C., Bailly, K., and Richard, G. (2017). Règles d'associations temporelles de signaux sociaux pour la synthèse de comportements d´agents conversationnels animés : application aux attitudes sociales. *Revue d'Intelligence Artificielle*

Ce qu'il faut retenir :

Question de recherche :

Comment utiliser des signaux sociaux extraits automatiquement pour modéliser une dynamique propre à l'expression de comportements socio-émotionnels donnés ?

Objectif :

- Intégrer l'information temporelle dans les modèles de phénomènes affectifs.
- Adapter cette information aux contraintes de la synthèse d'agent à partir de signaux extraits automatiquement
- Proposer une méthodologie d'évaluation dans le design de cette approche

Application :

Les attitudes sociales au sens d'Argyle, décrites selon deux axes :
Dominance et Appréciation

Chapitre 2

État de l'art

Sommaire

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 de données, généralement audio-visuelles, 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 agent avec l'expression d'attitudes crédibles. Ce chapitre propose de passer en revue les principales méthodes existantes et leurs conclusions selon deux grands axes.

Le premier axe s'intéresse au traitement des données à utiliser dans les modèles. Initialement, elles étaient obtenues manuellement : un annotateur décrit les gestes présents et note les tours de parole ou la présence d'un sourire par exemple. Ces observations humaines ont permis d'élaborer des descripteurs de plus en plus complexes. L'utilisation d'algorithmes en traitement du signal et de l'image permet maintenant d'estimer également ces caractéristiques automatiquement. Ces estimations des signaux sociaux extraits automatiquement ne nécessitent pas d'intervention humaine et sont donc plus faciles à obtenir. Cependant, les résultats obtenus peuvent présenter du bruit ou des erreurs et ne sont pas forcément toujours adaptés à une tâche de synthèse.

Le deuxième s'intéresse à la place de la temporalité dans les modèles. A partir des premiers modèles psychologiques ou sociologiques historiques, l'outil statistique va ensuite permettre de trouver des informations complémentaires lors de l'analyse des données. Les résultats obtenus ont servi de base pour la validation de modèles informatiques qui ont la capacité d'analyser des corpus de données de plus en plus importants. Ainsi, les algorithmes de fouille de séquences ou d'apprentissage profond vont apporter une information temporelle essentielle dans la construction de modèles d'expressions d'attitudes sociales. Leurs applications dans des tâches de synthèse devraient donc améliorer la génération de comportements d'agents conversationnels animés.

Une analyse des avancées pour chaque axe permet d'illustrer les principales méthodes utilisées par la communauté mais aussi d'identifier les informations manquantes. La suite de ce chapitre va donc passer en revue les grandes familles de méthodes possibles. Généralement, un ou plusieurs exemples concrets servent d'illustration afin de positionner les travaux de cette thèse et justifier ainsi les questions de recherche établies. Cela permet également d'obtenir des résultats validés par la littérature auxquels sont confrontés les études menées dans cette thèse.

2.1 Travail sur les signaux

Avant de pouvoir appliquer un modèle pour trouver des connaissances utiles à la reconnaissance ou la synthèse de comportement, il est nécessaire d'avoir des données sur lesquels travailler. La figure 1.3 montre que les données brutes sont généralement des fichiers audio-vidéo et qu'une étape d'extraction des signaux est nécessaire avant de nourrir le modèle. Différentes approches sont alors possibles.

Les signaux peuvent être extraits manuellement : des annotateurs, experts ou non, regardent les fichiers et vont indiquer les signaux présents à chaque moment.
Cette approche peut couvrir l'ensemble des données. Ainsi Chollet et al. (2013) utilisent l'outil d'annotation ELAN [1] pour repérer les mouvements de tête et les regards dans la conception d'un corpus d'analyse. Ce même outil est généralement utilisé pour que des observateurs extérieurs annotent les états affectifs présents dans l'interaction. C'est ce qui a été fait par exemple dans la conception du corpus *Semaine-db*(McKeown et al., 2012) où des annotateurs notent leurs impressions en terme d'expressions d'émotions, de tension, d'antagonisme. . . Ce procédé nécessite beaucoup de ressources humaines et est sensible à l'erreur humaine. Par exemple, le temps de réaction pour modifier un jugement peut être important.

1. https://tla.mpi.nl/tools/tla-tools/elan/

Une solution est d'utiliser des outils de traitement du signal avant de procéder à une vérification humaine pour contrôler les résultats. Bawden et al. (2015) dans leur analyse prosodique du corpus Semaine(McKeown et al., 2012), a annoté manuellement les actes de dialogue (assertifs, directifs, expressifs,...) mais les caractéristiques prosodiques ont été extraites avec le logiciel Praat [2](Boersma and Weenink, 2017) puis vérifiées manuellement. Cela permet un gain de temps conséquent mais nécessite toujours l'avis d'un expert.

Des solutions existent pour gommer les erreurs d'annotations. L'utilisation de l'outil statistique va ainsi pouvoir mesurer l'accord inter-annotateurs pour garantir la fiabilité d'un jugement.
Une autre solution consiste à utiliser des *clusters* pour regrouper les données en valeurs cohérentes. Ainsi, Cowie et al. (2010) proposent cette approche 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. Les mouvements étaient extraits manuellement et une analyse par *clusters* permet de les symboliser par les centroïdes des regroupements calculés. Ils trouvent alors 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.
Cette approche a également été explorée par Dermouche and Pelachaud (2016) où l'algorithme d'exploration de données intitulé *Apriori*, décrit dans Srikant and 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.

Certaines études évitent cette étape d'extraction de signaux. Pour cela, les intervenants sont équipés de capteurs. Bailly et al. (2015) ou Busso et al. (2008) utilisent des outils de capture de mouvements pour obtenir directement des données "bas niveau" grâce à des capteurs qui peuvent être placés sur les protagonistes des fichiers évalués. Cette approche intrusive présente le risque de modifier le comportement des intervenants à cause de la gêne qu'elle peut causer.

Ravenet et al. (2013) proposent une approche intéressante avec la création d'un corpus de postures d'agents conversationnels animés 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

2. http://praat.org/

soumise ou poser une question gentiment par exemple). L'intérêt de cette approche est d'utiliser des signaux sociaux provenant de l'agent et non d'un humain, ce qui est plus cohérent pour évaluer la perception du phénomène affectif présent pour une tâche de synthèse ensuite. Ravenet et al. (2013) 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.

Finalement, les progrès en traitement du signal et vision par ordinateur ont permis l'extraction automatique de données.
Dans Yu et al. (2017), les caractéristiques visuelles et acoustiques ont été extraites automatiquement avec les outils OpenFace(Baltrušaitis et al., 2016) et Covarep (Degottex et al., 2014). Ces travaux, contemporains de cette thèse, utilisent l'apprentissage profond pour prédire des états émotionnels grâce à un réseau LSTM.
En allant encore plus loin, Trigeorgis et al. (2016) intègrent l'extraction de caractéristiques directement dans le modèle. L'utilisation combinée de réseaux neuronaux convolutifs et de réseaux neuronaux récurrents permettent de prendre en entrée le sonagramme d'un signal audio et d'en établir l'émotion présente. Les performances ainsi obtenues dépassent les méthodes traditionnelles. Cependant, tout comme les règles extraites, les caractéristiques trouvées automatiquement par le modèle deviennent difficilement interprétables.

Le traitement de l'information, en prenant en compte des connaissances humaines a priori, permet également d'obtenir des résultats plus intéressants. Cela est particulièrement vrai pour l'apprentissage profond.
Ainsi Xu et al. (2016) proposent d'utiliser des LSTM bidirectionnels afin d'améliorer la synchronisation d'expressions faciales et des mouvements de lèvres en fonction de la parole tout en ajoutant une information émotionnelle. Ils explorent différentes stratégies pour améliorer la synthèse d'expressions faciales en incorporant de l'information issue d'un corpus neutre à l'information émotionnelle du corpus d'étude *INTERSPEECH 2009 Emotion Challenge*. Ils montrent ainsi des résultats encourageants sur l'utilisation de l'ajout d'information grâce à ces techniques pour la synthèse de phénomènes affectifs, en particulier en concaténant les prédictions issues des données des différents corpus.

2.2 Temporalité

Initialement, les recherches utilisent l'observation d'interactions pour en faire des analyses qualitatives ou quantitatives afin de déterminer les signaux ayant influencés la perception des attitudes exprimées par les intervenants. C'est cette méthode qui prévaut dans les analyses psycho-sociologiques présentées dans la partie 1.3.2.

Ainsi, Tusing and 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. Pour évaluer la dominance exprimée, ils demandent à des annotateurs de juger celle-ci. Un premier groupe annote le message même (dominance du contenu linguistique), un autre juge les vidéos sans le son (dominance du contenu visuel) et un dernier annote les vidéos. En parallèle, des données comme la fréquence fondamentale, l'intensité ou le débit sont extraites de chaque vidéo.
Les analyses se font donc au niveau de chaque extrait, permettant de relier certaines de ces caractéristiques à différentes variations de perception de dominance. Par exemple, Tusing et Dillard 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 féminines.
Dans un esprit similaire, 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 l'amicalité mais que le sourire influe principalement sur cette dimension. Ils ont cependant toujours utilisé la même dynamique temporelle entre les différentes situations.
Ces analyses sont intéressantes car elles pointent des contextes et des signaux influençant la perception de dominance. Cependant, leurs résultats n'offrent pas d'informations sur la dynamique de ces signaux : les caractéristiques sont à l'échelle du fichier étudié et perdent alors l'information sur l'évolution de l'interaction.

Des études prennent en compte cette information temporelle avec différents formalismes. Une première idée consiste à utiliser différentes fenêtres temporelles pour avoir une idée des dynamiques locales au sein d'un fichier global. Les deux exemples qui suivent, bien qu'orientés analyse et détection, pourraient donner des résultats intéressants pour reproduire des attitudes chez un agent virtuel.
Ward and Abu (2016) utilisent différentes fenêtres temporelles pour observer les variations de prosodie entre un expert et un novice de jeux vidéo. Cela permet d'avoir une information dynamique sur différentes échelles. Ils trouvent alors des co-occurences intéressantes entre les différentes phases du jeu et le statut de chacun des joueurs.
Une autre utilisation de fenêtres temporelles est proposée par Audibert (2007) avec une étude de la modélisation des expressions prosodiques des affects. Il utilise le principe de

gates afin d'observer l'impact de la temporalité : il s'agit de couper le stimulus original en des points prédéfinis et de le compléter avec du bruit blanc. Les fichiers audio originaux qui ont servi de stimulus étaient émotionnellement marqués et cette découpe en *gates* donne une indication sur l'influence de la temporalité dans la perception des émotions. Audibert analyse la perception de l'état émotionnel dans ces stimuli audio dont le contour prosodique est contrôlé. 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 à d'autres états émotionnels.

Des informations contextuelles sont aussi observées, en particulier pour trouver des informations sur la synthèse d'un locuteur.
Ainsi, 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 ou la trame d'une vidéo.
Dans un objectif de synthèse, Bawden et al. (2015) effectuent une analyse prosodique du corpus Semaine (McKeown et al. (2012), détaillé dans la section 3.3.2). Ils explorent les relations entre la personnalité, le type d'acte de dialogue (assertifs, directifs, expressifs,...) et des caractéristiques prosodiques. L'analyse prosodique montre 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 des 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.

Ces approches statistiques donnent ainsi des indications sur les signaux pertinents. Elles fournissent également des premiers éléments soulignant l'intérêt d'étudier la temporalité mais les résultats restent trop limités pour des tâches de synthèse.
Par ailleurs, des algorithmes d'apprentissage automatique sont de plus en plus utilisés pour chercher de l'information dynamique permettant la synthèse.
Ainsi, Lee and Marsella (2012) ont étudié trois algorithmes d'apprentissage (modèles de Markov cachés (*HMM*), champs aléatoires conditionnels (*CRF*) et des champs aléatoires conditionnels dynamiques à variables latentes (*Latent-Dynamic Conditional Random Fields*, *LD-CRF*)) pour modéliser l'amplitude des mouvements de tête et de sourcils d'un orateur. Les *HMM* sont des modèles statistiques très répandus pour des problèmes où la dynamique est importante (synthèse vocale, reconnaissance d'écriture, ...). Les *CRF* permettent de repérer des dépendances sur des temps plus long : Morency et al. (2008) l'illustrent sur de

la prédiction de mouvements de tête tout comme les *LD-CRF*.
Lors de l'évaluation de ces solutions, Lee and Marsella (2012) montrent les bonnes performances du *LD-CRF* pour prédire la coordination des mouvements de tête et de sourcils. Cependant, une étude perceptive n'a pas montré de différences significatives entre leur modèle et celui de la littérature. Elle prouve néanmoins la faisabilité et l'intérêt d'un modèle construit automatiquement prenant en compte la dynamique des signaux.

Une autre solution pour faire de la génération d'agents consiste à rechercher des motifs utilisables en entrée d'algorithmes d'apprentissage automatique. La fouille de données (plus détaillée dans la partie 4.1) a déjà montré des résultats intéressants dans des systèmes de création de dialogue pour des agents conversationnels animés (Ales et al., 2012). Cette approche est également explorée par Martínez and Yannakakis (2011) puis Chollet et al. (2014) qui proposent d'utiliser des algorithmes de fouille de données pour trouver des séquences simples de signaux non-verbaux associées à des attitudes sociales.
Ainsi Martínez and Yannakakis (2011) se placent dans le contexte des jeux vidéo pour relier des données à des émotions comme la frustration. Ils utilisent l'algorithme *Generalised Sequence Pattern* (GSP), décrit dans Srikant and 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 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, si GSP trouve des séquences d'événements, 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 agent virtuel 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.
L'utilisation d'algorithmes de fouille de données a aussi été explorée 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.

L'apprentissage profond a récemment permis l'étude de la dynamique des signaux grâce à l'apparition des réseaux neuronaux récurrents et des LSTM (Long-short Term Memory, voir partie 5.1.2) en particulier.
On peut observer cette évolution à partir des travaux de Ding et al. (2015a). L'utilisation de réseaux neuronaux ont montré leur efficacité pour effectuer de la synthèse de mouvements de tête à partir d'un signal audio, améliorant les résultats des réseaux de Markov cachés. L'utilisation de LSTM bidirectionnels (Ding et al., 2015b; Greenwood et al., 2017) et de techniques d'engorgements ("bottleneck") (Haag and Shimodaira, 2016; Lan et al.,

2016) ont ainsi amélioré le réalisme de la synthèse de ces mouvements à partir du signal audio. Avec le même type de techniques, la synthèse de mouvements de lèvres a été étudiée avec de bons résultats. Suwajanakorn et al. (2017) en particulier proposent de modifier directement dans des vidéos les mouvements de lèvres d'un protagoniste pour qu'ils correspondent au son voulu. Ils utilisent une décomposition pyramidale des vidéos selon les différentes parties du visage (mâchoires, cou, bouches,...) afin d'améliorer leurs résultats. Ils montrent l'efficacité de l'ajout de ces connaissances en testant leur approche sur des vidéos d'allocution du président Obama.
Pour de la synthèse d'agents, Sadoughi et al. (2017) ont proposé une architecture jointe pour effectuer la synthèse de différentes zones du visage en fonction de la parole. A partir de cepstre, un modèle LSTM bidirectionnel va apprendre, d'abord conjointement puis en parallèle, comment animer le haut, le milieu et le bas du visage. Appliqué au corpus *IEMOCAP*, des améliorations objectives ont été trouvées, mais les études perceptives ne soulignent pas d'améliorations par rapport aux références. Les auteurs indiquent que l'ajout d'informations pourrait mener à de meilleurs résultats.
Ces modèles sont efficaces pour trouver de l'information temporelle dans la dynamique des signaux et prédire les évolutions de ces derniers, mais leurs résultats et les représentations trouvées ne sont pas évidents à interpréter pour des humains. Par ailleurs, ces approches ont besoin de beaucoup de données afin d'avoir des résultats pertinents (comme cela sera détaillé dans la section 5.1).

Ces différentes approches montrent comment l'information temporelle a été progressivement intégrée dans les modèles. Il ressort néanmoins qu'il faut trouver un compromis entre les performances des modèles et l'interprétabilité des résultats dans la plupart des solutions étudiées.

2.3 Positionnement

UNE information temporelle précise reste donc souvent la partie manquante de ces solutions pour générer efficacement des agents conversationnels animés capables 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 de la durée d'un sourire. Les précédentes études font cependant ressortir les signaux qui influencent la perception d'attitudes comme cela est résumé dans le tableau 2.1.

Dans cette thèse, deux approches ont été explorées pour effectuer cette analyse de la dynamique des signaux sociaux. Par rapport aux travaux existants, le but est de proposer l'approche la plus automatique possible. Les signaux sociaux sont bas niveaux et extraits automatiquement des données d'études. Les modèles recherchent l'information tempo-

Modalité	Références	Influence la dominance	Influence l'appréciation
Prosodie	Tusing and Dillard (2000)	l'énergie de la voix, ses variations et le débit. F0 moyenne pour les hommes	-
	Vinciarelli et al. (2009b)	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 et al. (2010)	-	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. (2009b)	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

TABLEAU 2.1 – 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.

relle en exploitant le plus possible toute information, même vague, afin de la lier à un état affectif voulu. Ces modèles doivent ensuite permettre simplement de générer le comportement voulu.

La première approche, *SMART*, détaillée dans la section 4, propose l'utilisation de règles d'associations temporelles modélisant les liens entre divers signaux lors de l'expression d'attitudes sociales. Ce système est basé sur l'algorithme de fouille de séquences *TITARL* de Guillame-Bert and 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.
La méthode *SMART* encapsule et adapte *TITARL* dans une architecture permettant d'analyser des signaux bas niveaux, extraits automatiquement de flux audio-vidéos et donc potentiellement bruités. La fouille de séquences de *TITARL* permet d'en extraire des règles d'associations temporelles pertinentes et précises. *SMART* permet de les calculer pour qu'elles

soient caractéristiques d'un phénomène affectif. *SMART* les lie donc à l'expression d'attitudes et va les enrichir pour permettre la synthèse du comportement d'un ACA exprimant l'attitude désirée.

La deuxième approche, le *Social Separation Network* (*SSN*), section 5, se focalise plus sur l'influence d'une expertise humaine dans la construction d'un modèle d'étude automatique de la dynamique de phénomènes affectifs. *SSN* utilise l'apprentissage profond et la séparation de domaine afin de trouver automatiquement des représentations dans la dynamique de signaux sociaux liés à l'expression d'états affectifs.
Ce réseau souligne ainsi comment l'utilisation d'indicateurs humains pour améliorer une architecture d'apprentissage profond modifie les résultats, même avec un faible jeu de données. Cette approche n'a pas encore donné lieu à une synthèse mais la méthode pour y aboutir est détaillée.

Ainsi, les approches proposées dans ce manuscrit veulent analyser des signaux sociaux bas niveaux, extraits automatiquement, qui permettent d'exprimer dans le temps un ou plusieurs états affectifs donnant ainsi des indications pour une tâche de synthèse d'un agent conversationnel animé. Les deux modèles proposés illustrent aussi comment l'utilisation d'information a-priori peuvent améliorer les résultats.

Ce qu'il faut retenir :

Positionnement :

- recherche d'informations temporelles précises liées à l'expression d'attitudes
- utilisation de signaux extraits automatiquement, relativement bas niveau et potentiellement bruités
- analyse des résultats de la littérature afin d'avoir une information a-priori pour valider les résultats des études effectuées lors de cette thèse

Chapitre 3

L'analyse de corpus

Sommaire

Ce chapitre présente les différents corpus de travail qui ont été étudiés dans cette thèse. Une analyse de l'existant justifie les choix d'études et les limitations pour l'application à la synthèse d'attitudes chez un agent conversationnel animé. Il introduit également le corpus *President Of The United States* (*POTUS*) qui a été construit pour compléter ces études. Il est composé de vidéos d'allocutions du président Obama également reproduites avec un agent conversationnel animé afin d'étudier les différences selon l'intervenant.

3.1 Introduction

L'analyse de corpus est la méthode la plus courante pour construire un modèle en vue de la synthèse de comportement chez un agent. Dans les travaux de cette thèse, l'approche classique a été employée, telle que Cassell (2007) l'a formalisée dans son cycle de développement *Study-Model-Build-Test*. Cette méthodologie circulaire, illustrée dans la figure.3.1, débute par la collecte de données. Leur interprétation va permettre d'élaborer un modèle formel qui sera implémenté sur une plate-forme d'agents virtuels. Ce résultat sera évalué par des humains afin de le corriger voir d'initier une nouvelle collecte de données si des manquements trop importants sont remarqués.

Les corpus sont généralement composés d'un ensemble de fichiers, souvent multimédia, contenant des interactions d'humains (parfois d'agents ou parfois d'humains seuls) qui sont annotés afin de décrire le comportement communicatif présent. Cette méthodologie vient des sciences sociales qui ont développé ce procédé. Elle a permis l'élaboration d'une littérature riche, basée sur l'observation, qui a été présentée dans la section 1.3.
Rehm and André (2008) nuancent cette richesse en soulignant que les corpus et modèles alors proposés ne prenaient pas en compte les contraintes liées à la génération, telles que

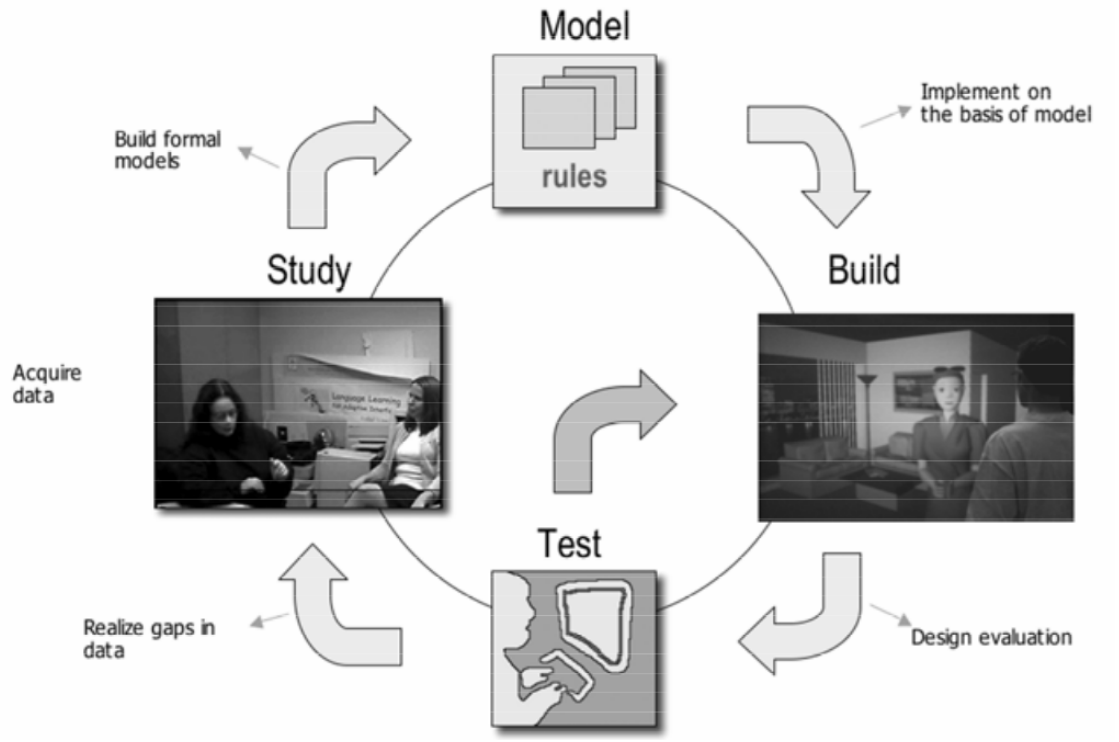

FIGURE 3.1 – Méthodologie pour la construction de modèles de synthèse pour des agents virtuels, issue de Cassell (2007)

la multimodalité ou la synchronisation des signaux. C'est pour cela que de nouveaux corpus ont dû être construits pour la synthèse d'agents conversationnels. Par ailleurs, Rehm and André (2008) indiquent que deux stratégies peuvent être appliquées pour la synthèse à base de corpus : soit en clonant directement des comportements issus du corpus correspondant à l'affect désiré, soit en extrayant des règles qui construiront un modèle commun pour la génération. Ce point sera observé dans l'élaboration du *POTUS Corpus*.

Les travaux de cette thèse se sont limités aux interactions face à face, le but étant de proposer des approches nécessitants le moins d'interventions humaines possibles. Dans la suite de ce chapitre, les algorithmes et traitement utilisés pour extraire les signaux sociaux de vidéos vont être détaillés. Cette approche automatique implique un certain nombre de contraintes sur les corpus de travail, et en particulier sur la qualité des données pour assurer des résultats corrects en traitement de la voix et vision par ordinateur. Une revue de corpus existants illustrera et justifiera les études menées. Enfin, le *POTUS Corpus* sera présenté et une analyse statistique le détaillera.

3.2 L'extraction des signaux

Les travaux de cette thèse s'intéressent aux signaux liés au visage et à la parole. Comme expliqué dans la section 1.4, différents algorithmes ont été utilisés pour estimer ces signaux dans des vidéos. Pour cela, un ensemble de signaux sociaux d'étude a été défini, composé *d'informations prosodiques*, *des activations des action units*, *les mouvements de tête* et d'informations telles que *les tours de parole*. Les caractéristiques sélectionnées vont maintenant être décrites tout comme les méthodes pour les évaluer.

Les tours de paroles indiquent si l'intervenant est en train d'écouter ou de parler ainsi que les moments de prise et de fin de parole. Ces informations peuvent provenir des transcriptions fournies par le corpus étudié. Cependant, un outil de détection automatique de tour de parole a également été utilisé (De Jong and Wempe, 2009). Il s'agit d'un script praat [1] capable de détecter des silences selon l'intensité acoustique du signal et une durée minimale.
Ces données peuvent aussi être utilisées pour ajouter de l'information à d'autres signaux en les enrichissant avec des informations contextuelles, par exemple indiquer pour les *action units* si la personne est orateur ou auditeur.

Les descripteurs prosodiques ont été extraits avec des solutions présentant des approches différentes.
La première, Prosogram, est un programme développé par Mertens (2004) qui propose

1. http://www.praat.org/

une segmentation automatique des fichiers audio en syllabes notées *nucleï* puis calcule des paramètres prosodiques globaux tels que la gamme de pitch de l'orateur. Il transforme ensuite cela en approximation des mouvements de pitch perçus. Cette approche "*bottom-up*" a l'avantage de ne pas avoir besoin d'informations supplémentaires (annotations, entraînements, ...) et limite donc le risque de biais. Pour chaque *nucleï* la fréquence fondamentale moyenne (f_0), ses variations, les pics d'intensité et la forme du pitch (montée, descente, plat, ...). Les formes du pitch sont ensuite concaténées grâce aux transcriptions pour avoir cette information au niveau du mot avec son timing. Les trois autres données étant continues (f_0, ses variations et les pics d'intensité), celles-ci sont ensuite transformées en événements symboliques.
La seconde, COVAREP (Degottex et al., 2014), fournit une boite à outils Matlab qui permet de calculer plusieurs descripteurs de la prosodie et de la qualité de la voix. Il s'agit d'un ensemble d'algorithmes de traitement de la parole validés, testés et largement employés dans la communauté (Valstar et al., 2016). Cette thèse utilise la fréquence fondamentale calculée toutes les 0.01 secondes.

Les *action units* sont des estimations des mouvements du visage selon le *facial action coding system* introduit par Ekman and Friesen (1978) [2], illustrées dans la figure 3.2. Les *action units* ont été automatiquement extraites grâce à la solution de Nicolle et al. (2016). Afin de réduire le bruit de cette détection automatique, un lissage exponentiel ("*exponential smoothing*") a été appliqué avec un coefficient de lissage α égal à 0.5. Les *action units* sont ensuite symbolisées selon trois cas possibles : désactivée, faible activation et forte activation. Les études présentées ensuite se limitent aux *action units* relatives aux sourcils (AU_1 et AU_2 regroupées, pour le haussement, AU_4 pour le froncement), aux pommettes (AU_6) et aux coins des lèvres (AU_{12} pour la hausse des commissures) .

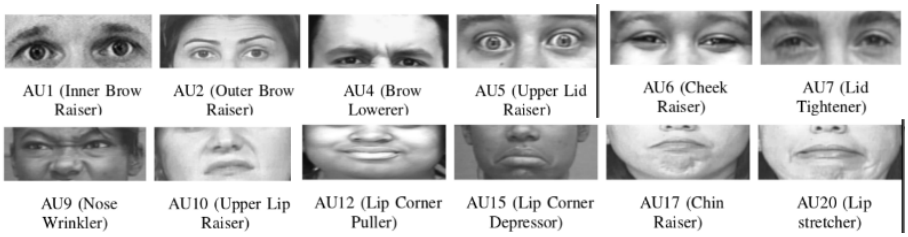

FIGURE 3.2 – Facial Action Unit correspondant à l'activation de différents muscles faciaux. Images obtenues via http://www.cs.cmu.edu/~face/facs.htm

2. http://www.la-communication-non-verbale.com/2013/03/facial-action-coding-system-6734.html

3.3 Présentation des corpus de travail existants

3.3.1 Revue de corpus

De nombreux corpus existent pour étudier différents phénomènes affectifs en proposant des supports ou des modalités variés ainsi que divers jeux d'annotations, plus ou moins riches ou précis. Pour rappel, le cadre d'application de cette thèse consiste à la synthèse d'attitudes chez un agent virtuel lors d'une interaction face-à-face. Une première étape dans cette thèse a été d'étudier les bases de données existantes afin de déterminer celles qui peuvent satisfaire les contraintes établies. Les méthodes proposées nécessitent des fichiers audio et vidéos de bonne qualité afin de pouvoir extraire automatiquement les caractéristiques (détaillées dans la section 3.2), si possible avec des annotations en attitude (ou pouvant être annotés).

La contrainte principale tient dans le fait qu'une attitude a besoin de temps pour se développer. Ainsi, *Gemep* (Bänziger and Scherer, 2010) ou *AFEW-VA* (Kossaifi et al., 2017) proposent des séquences trop courtes pour les études menées. Le deuxième point bloquant concerne l'extraction des *action units*. La solution de Nicolle et al. (2016) requiert une vidéo de bonne qualité où le visage est bien visible et de face. Ainsi, des corpus comme le *Canal 9 political debate* (Vinciarelli et al., 2009a) ou le *TED database* (Pappas and Popescu-Belis, 2013) ont de nombreux changements de plans rendant l'algorithme inefficace. Des corpus ont été également filmés de biais comme *IEMOCAP* (Busso et al., 2008) ou le corpus *Tardis* (Chollet et al., 2013) ce qui les disqualifie également, tout comme *Recola* (Ringeval et al., 2013) où les participants, lors d'une tâche de survie, regardent une feuille en penchant la tête ce qui affecte les performances du détecteur.

3.3.2 Semaine-db

Un compromis a été trouvé avec l'étude de 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. Le coté "émotionnellement colorées" signifie ici que les acteurs vont surjouer leurs personnages pour exprimer de façon caricaturale de l'hostilité, de la joie ou de la dépression.
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. La base propose des vidéos sonorisées de bonne qualité des visages des participants et des utilisateurs que les algorithmes d'extraction de caractéristiques peuvent analyser efficacement.

Dans les études menées, seule la partie opérateur a été considérée : à chaque session, l'acteur joue quatre rôles prédéfinis. Spike est agressif, Poppy est amical, Obadiah est dépressif et Prudence pragmatique. Seuls les rôles de Poppy le gentil et Spike le méchant ont été retenus pour les comparer, car ils se situent aux extrêmes de l'axe de l'amicalité sur le circomplexe d'Argyle. Cela représente 25 enregistrements de Poppy et 23 de Spike de 3-4 minutes chacun, joués par quatre acteurs différents, soit environ 150 minutes de données. 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.

3.4 Le corpus *POTUS*

3.4.1 Introduction et motivation

Les limites des corpus existants, liées à la qualité des vidéos et aux annotations disponibles, ont motivé la création d'un corpus de travail dans cette thèse. Afin de trouver des informations intéressantes pour la synthèse du comportement d'un agent virtuel, des vidéos de bonnes qualités dans lesquelles les protagonistes exagèrent l'expression de leur état émotionnel ont été recherchées. Par ailleurs, la contrainte de détection automatique des *action units* nécessite que l'intervenant soit face caméra.

Ces contraintes ont rapidement orienté les recherches des fichiers audio-vidéos nécessaires pour la réalisation du corpus à des allocutions de personnalités politiques. Des études précédentes, dont D'Errico et al. (2013), confortent ce choix, car ces individus sont généralement bien formés au contrôle de leurs images pour toucher un public bien précis. Les allocutions face caméra récentes permettent ainsi d'avoir une extraction des signaux sociaux satisfaisante avec des messages variés. Une première étape dans la réalisation du corpus *POTUS* consiste donc en un jeu d'allocutions du président Obama lors de ses deux mandats.

3.4.2 Premier intervenant : président Obama

Durant ses mandats, le président Obama faisait chaque samedi matin une allocution au peuple américain appelée *weekly address* [4]. Dans chaque vidéo, le président Obama fait une courte allocution, généralement seul face caméra, pour commenter l'actualité et discuter de son action. Les talents de communicant de l'ancien président [5] rendent ces vidéos

3. https://youtu.be/Tt8UOw4-Mdw
4. https://obamawhitehouse.archives.gov/briefing-room/weekly-address
5. http://www.scienceofpeople.com/2015/02/body-language-leaders-president-obama/
https://www.fastcompany.com/1070311/communicative-power-barack-obama-how-he-became-president-elect
http://www.mediate.com/articles/sharlandA8.cfm

particulièrement intéressantes à étudier.

Pour la réalisation du *POTUS Corpus*, les allocutions de Thanksgiving et Pâques entre 2009 et 2015 ont été récupérées sur internet et analysées. Cela permet de valider l'approche retenue pour la construction de cette base de données qui pourra ensuite être enrichie d'autres allocutions. Les allocutions de Pâques et Thanksgiving ont été retenues car elles suivent globalement le même déroulé lié au contexte de ces festivités : elles commencent par un salut amical avant de traiter des sujets courants et finir sur Obama souhaitant de bonnes fêtes. Elles sont donc plus propices à des variations d'attitudes.
Par la suite, d'autres allocutions, sur d'autres périodes ou avec d'autres président, pourront être ajoutées pour compléter ce corpus.

3.4.3 Deuxième intervenant : agent Rodrigue

Pour chaque *weekly address* du président Obama, une vidéo similaire avec un agent virtuel a été réalisée qui sera intitulée Agent-Miroir par la suite. Pour cela, la plate-forme Greta a été utilisée avec l'agent Rodrigue. Le design de l'Agent-Miroir a été fait comme suit.

Premièrement, l'audio de la vidéo originale a été extrait. A partir de la vidéo, les mouvements de tête et les action units ont été évalués respectivement grâce aux algorithmes de Xiong and De la Torre (2013) (Intraface) et de Nicolle et al. (2016). Cependant, ces données extraites automatiquement présentaient du bruit et leur utilisation directement sur l'agent n'était manifestement pas satisfaisante. En effet, il pouvait y avoir des soubresauts non-naturels dans les *action units* ou les mouvements de têtes ce qui affectait le réalisme du rendu. Une opération pour trouver des paramètres correctifs d'ajustements a donc été mise en place pour assurer que les signaux affichés par l'agent et ceux détectés par les algorithmes soient cohérents.

Cette étape de calcul a utilisé des données artificielles : des vidéos de contrôle où l'agent utilisait une ou plusieurs de ses *action units* ou effectuait des mouvements de tête précis ont été générées. Les algorithmes ont ensuite analysé les vidéos pour estimer les intensités des *action units* et des mouvements de têtes. Les valeurs théoriques et mesurées ont ensuite été comparées. Il s'agit d'évaluer les paramètres correctifs sur différents sous-ensembles des valeurs théoriques et observées afin d'atteindre un consensus acceptable résistant au bruit de mesure.
Pour cela, une méthode itérative a été appliquée afin de trouver les paramètres de lissage et de transformation affine à appliquer entre le $signal_{\text{original}}$ et le $signal_{\text{extrait}}$. Un lissage local est appliqué au $signal_{extrait}$ sur une fenêtre de taille f, donnant le $signal_{\text{lissé}}$. Plusieurs fenêtres temporelles sont ensuite choisies aléatoirement. Sur chacune de ces fenêtres, les

moyennes et variances du $signal_{\text{original}}$ et du $signal_{\text{lissé}}$ sont comparées afin de trouver les coefficients de recalage. Ces coefficients sont ensuite appliqués à l'intégralité du $signal_{\text{lissé}}$ et l'erreur avec le $signal_{\text{original}}$ est calculée. Si l'erreur est inférieure à l'erreur précédemment calculée, les coefficients sont conservés. Ce processus est répété plusieurs fois pour trouver les coefficients optimaux. Les différentes étapes sont illustrées dans la figure 3.3. Cette méthode a ainsi été appliquée pour corriger les valeurs des *action units* et des mouvements de têtes extraits automatiquement. Une vidéo de ce procédé est visible ici [6].

Pour finir le comportement de l'agent miroir, une translation de sa tête équivalente au mouvement original du président Obama a également été ajoutée ainsi que des clignements de paupières aléatoires selon le processus de la plate-forme Greta. Enfin, la voix d'Obama a été ajoutée à la vidéo de l'Agent-Miroir pour obtenir le résultat visible ici [7].

3.4.4 Annotations

Suivant les travaux de (Cafaro et al., 2016; Zhao et al., 2016), des tranches de 15 secondes (15s *thin slices*) ont été utilisées pour évaluer les attitudes sociales exprimées dans les vidéos. Le principe est de découper chaque vidéo en segments de 15 secondes qui seront annotés indépendamment et aléatoirement. Cette technique du *thin slice*, introduite par Ambady and Rosenthal (1992), a été étudiée et validée par Murphy (2005) comme permettant d'avoir des jugements cohérents sur une interaction plus longue. Cela permet d'observer l'évolution des attitudes sur des pas de temps relativement fins.
Un premier test d'annotations à été effectué afin de valider l'intérêt de la construction de ce corpus. Pour cela, le public de la Cité des Sciences, un musée dont le but est la diffusion de la culture scientifique, a jugé les vidéos lors d'ateliers. Des différences notables dans

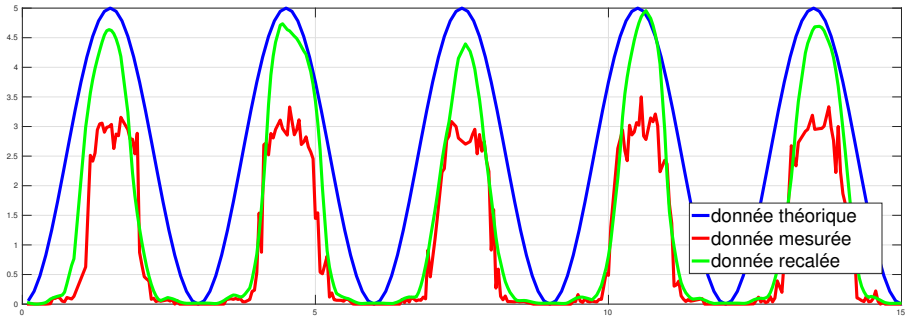

FIGURE 3.3 – Les différentes étapes de l'opération de recalage entre le signal théorique, le signal extrait automatiquement (mesuré) et le signal recalé.

6. https://youtu.be/C83s3wvyUF0
7. https://youtu.be/Yf7Ze7nQNbw

les jugements des visiteurs ont permis de valider cette approche et motivé la campagne d'annotation.
Ensuite, la plate-forme *Crowdflower*[8] a été utilisée pour demander à des anglophones de juger ces segments selon une douzaine de dimensions. Elle permet d'obtenir rapidement un grand nombre d'annotations. Les dimensions utilisées correspondent au questionnaire d'évaluation IPQ-r de Trapnell[9]. Les douze dimensions, aussi appelées duodecants, permettent une évaluation fine des attitudes selon les deux dimensions amicalité et dominance. Chaque duodecant était évalué grâce à une échelle de Likert de dimension 5 par cinq annotateurs différents. Les deux opérations de transposition, présentées dans les formules 1.1 et 1.2, permettent d'évaluer ensuite les attitudes selon les deux axes voulus comme cela est illustré dans l'image 3.4. Cette méthode permet d'éviter un biais lié aux sens des mots employés lors de l'évaluation et assure ainsi des résultats plus robustes.

3.4.5 Analyse

L'analyse présentée ici permet déjà d'observer la différence entre la perception des attitudes selon l'observation du Président Obama ou de l'Agent Miroir. Elle est motivée par le modèle cognitif de Brunswik (1956) (cf figure 1.2) qui souligne l'importance de l'externalisation et de l'attribution des signaux sociaux lors d'une interaction. En effet,

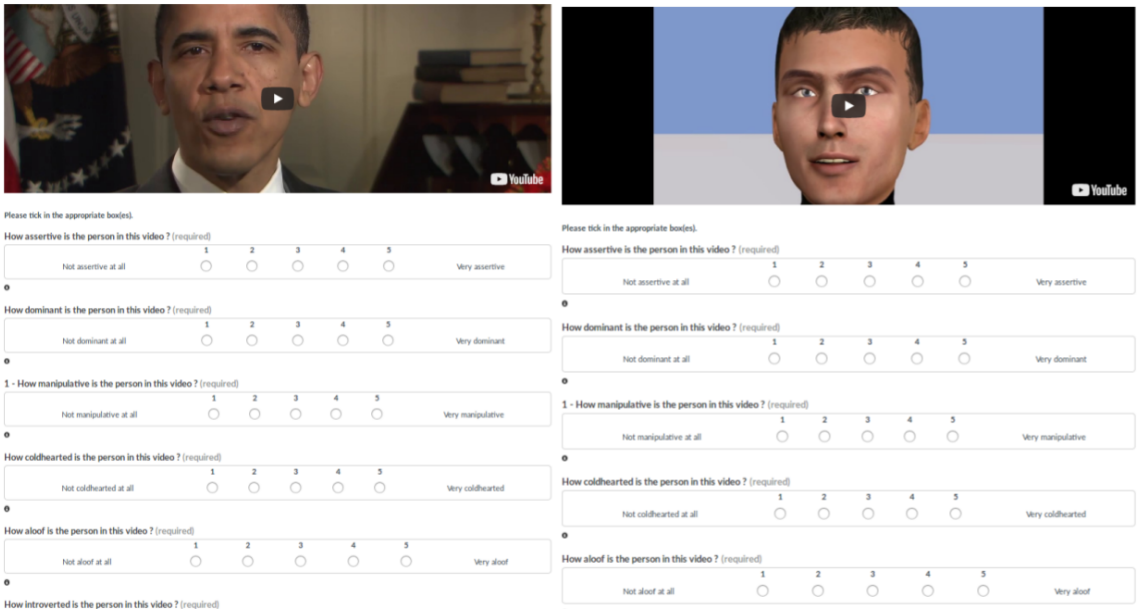

FIGURE 3.4 – Deux captures d'écran de la plate-forme *crowdflower* lors de la tâche d'annotation suivant le questionnaire de Trapnell.

8. https://www.crowdflower.com/
9. www.paultrapnell.com/measures/IPQ-revised.doc

l'Agent-Miroir est animé à partir des vidéos du président Obama et le copie au mieux. Cependant, l'approximation des valeurs des signaux extraits ainsi que le panel de signaux disponibles font que l'externalisation est différente. De plus, l'attribution d'un jugement par un observateur entre un agent et un ancien président a de fortes chances d'être différente. Cette étude cherche donc à quantifier ces différences afin d'évaluer l'influence de la personnalité dans la perception des attitudes sociales. Les deux cas étudiés ici sont 1) le locuteur est connu et humain, 2) le locuteur est inconnu et un agent virtuel.

Les annotations regroupées pour chaque *weekly address* sont visibles sur la figure 3.5. D'un point de vue global, la comparaison des moyennes des jugements montre de légères différences sur la perception de l'amicalité (1.72 en moyenne pour Obama contre 1.59 pour l'Agent-Miroir) et de la dominance (1.38 contre 1.25). Cependant, une analyse

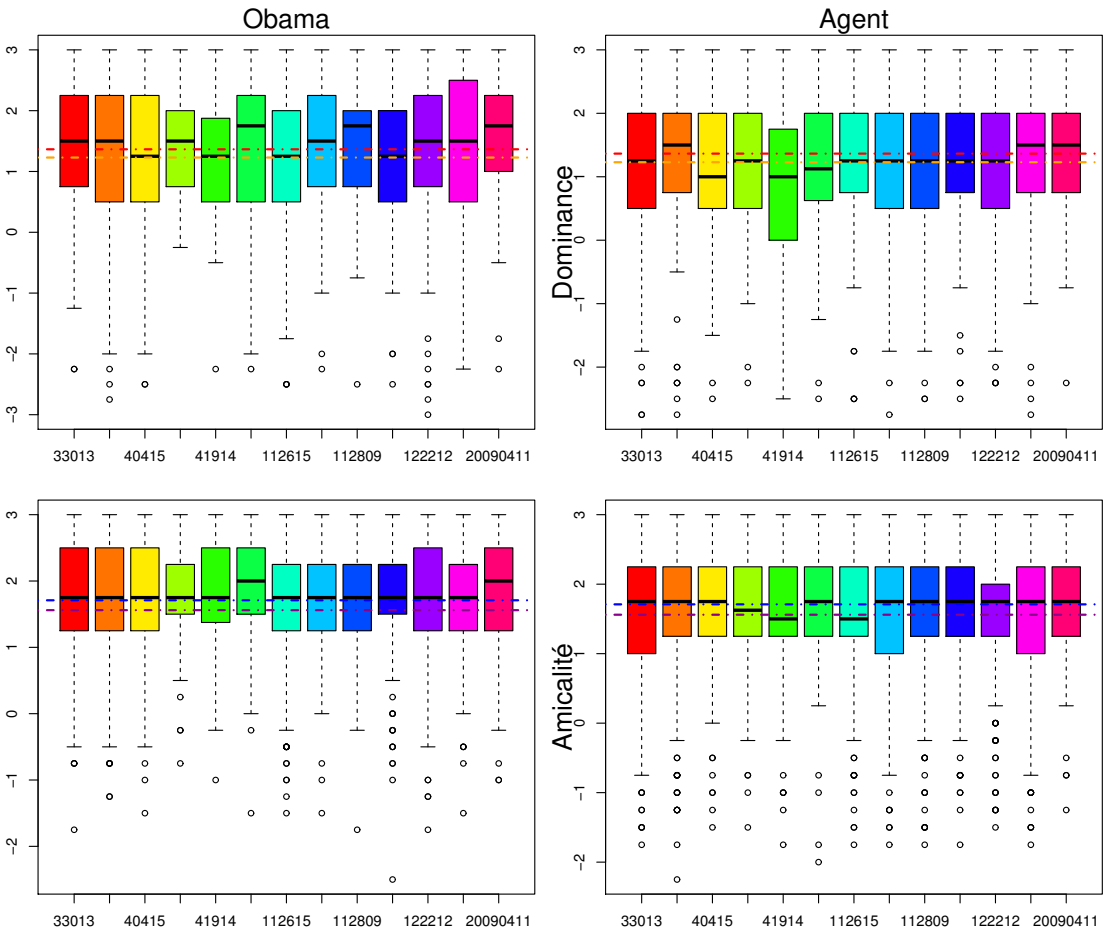

FIGURE 3.5 – Visualisation des annotations pour chaque *weekly address*. La dominance est en haut, l'amicalité en bas, Pr. Obama à gauche, l'Agent-Miroir à droite. Les lignes en pointillé indiquent les valeurs moyennes pour chaque annotation (rouge et bleu pour Obama, orange et violet pour l'agent). Chaque couleur correspond à un *weekly address*.

avec un test de Wilcox appareillé (Motulsky, 2013) montre que ces différences ne sont pas significatives. Cela indique que, globalement, la personnalité et l'incarnation du langage non-verbal ne semblent pas influencer la perception de la dominance et de l'amicalité.

Une analyse plus fine de l'annotation de chaque morceau de vidéo, chaque *slice* pour reprendre la terminologie de Zhao et al. (2016), permet de visualiser les différences dans la dynamique des attitudes perçues. Cela est visible dans la figure 3.6 qui présente l'évolution des annotations en dominance et en amicalité, pour chaque *weekly address*, au fil des slices. Cela permet d'observer localement les différences entre Barack Obama et l'Agent-Miroir.

Généralement, les annotations restent cohérentes entre l'humain et l'agent. Des écarts

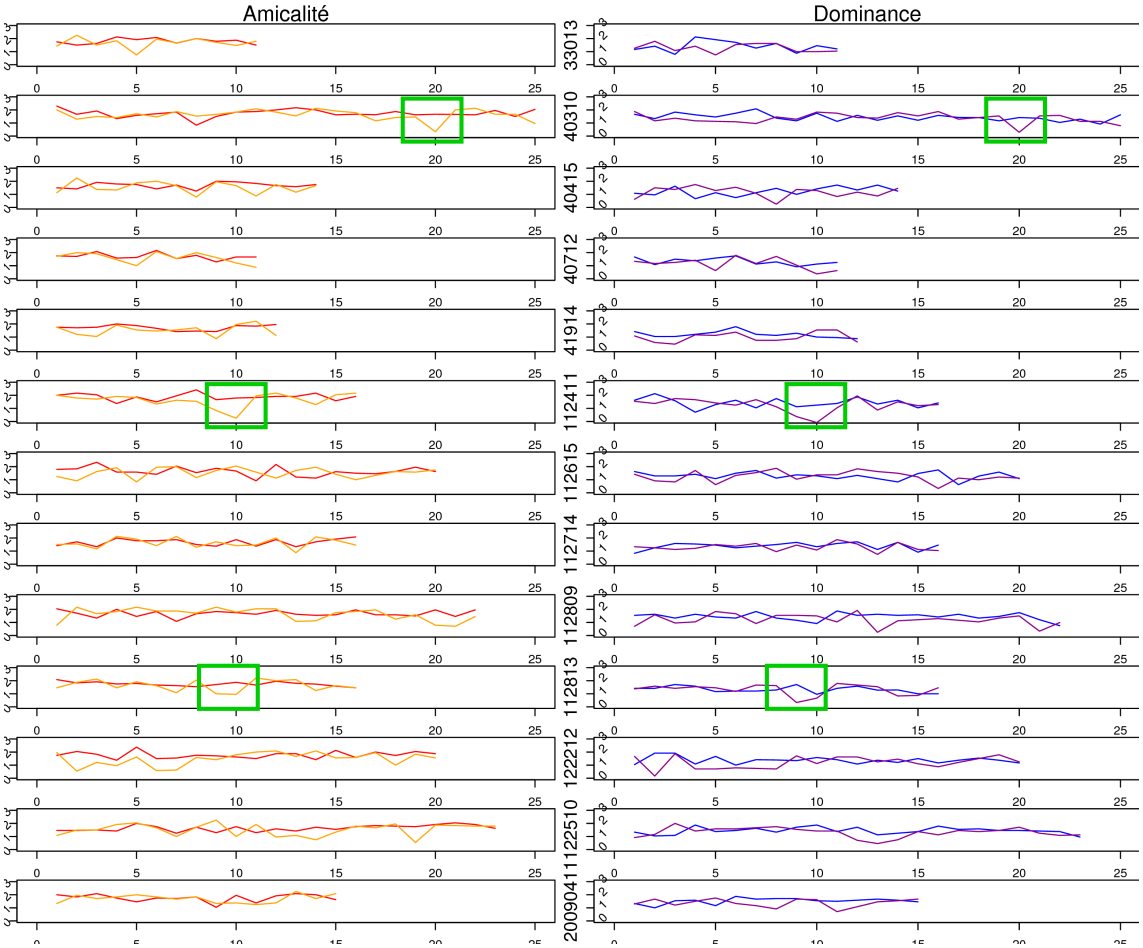

FIGURE 3.6 – Dynamique des annotations pour chaque vidéo, le temps est compté en *thin slice* d'annotation.
Légende :
rouge : amicalité Obama, orange : amicalité Agent,
bleu : dominance Obama, violet : dominance Agent.

peuvent être observés à certains endroits, les plus importants étant en amicalité et en dominance au même moment et toujours dans le même sens (baisse nette de la perception de l'agent par rapport à Obama, encadrés en vert dans la figure). Une piste d'explication après un visionnage des scènes en question peut résider dans une perception dérangeante de l'agent lié aux problèmes de vallée de l'étrange (Mori et al. (2012), *uncanny valley* en anglais). Il s'agit du phénomène selon lequel plus un agent ressemble à un humain, plus ses imperfections sont monstrueuses, donnant une sensation dérangeante voir morbide. Ce phénomène est donc encore en cours d'analyse.

Une analyse statistique a été menée au niveau des *slices* pour quantifier la différence entre les vidéos du Président Obama et celles de l'Agent-Miroir. Suivant les recommandations de Motulsky (2013), un test des rangs appliqué aux échantillons appariés (Wilcoxon matched-pairs) a été effectué. Il n'a pas montré de différence significative en ce qui concerne l'amicalité, mais un effet a été trouvé pour la dominance (p-value$=0.004 < 0.01$). Cela peut être expliqué par la stature de Barack Obama, ancien président internationalement connu, là où l'Agent-Miroir est un inconnu (pour l'instant).

3.5 Conclusion

Ce chapitre souligne les contraintes spécifiques à la synthèse de comportements d'agents qui s'appliquent à l'étude de corpus avec une extraction automatique de signaux sociaux (cf section 3.2). Généralement, le cas d'application voulu oriente le choix du corpus d'étude. Dans le cadre de cette thèse, la contrainte principale est de pouvoir observer des attitudes (soit des différences en amicalité ou en dominance). L'utilisation d'algorithmes automatiques pour extraire les signaux sociaux restreint considérablement le panel de corpus possible (cf section 3.3.1) car ils nécessitent des vidéos de bonne qualité dans lesquelles le visage de l'intervenant est bien visible.

Un consensus acceptable est trouvé avec le corpus *SAL-SOLID SEMAINE* : il propose des vidéos avec des expressions d'attitudes marquées et une qualité correcte pour l'extraction des signaux. Ces données permettent les premières études pour valider les modèles qui sont présentés dans les deux chapitres suivants.
Cependant, l'absence d'annotations fines en attitudes ainsi que la qualité juste acceptable des vidéos pour l'estimation des Action Units motivent la création du corpus *POTUS*. Il propose ainsi des vidéos sonorisées de très bonne qualité d'allocutions du Président Obama au peuple américain annotées en attitudes sociales. A cela s'ajoute un ensemble de signaux sociaux qui ont été extraits automatiquement, corrigés et ajustés afin de générer les vidéos d'un Agent-Miroir qui copie l'ancien président (cf section 3.4.3).

Les annotations de ces vidéos d'Agent-Miroir montrent que la perception d'attitudes reste similaire à celle des vidéos originales d'Obama en amicalité et légèrement modifiées en dominance (cf section 3.4.5).
Dans un premier temps, cela permet de valider les traitements des signaux extraits automatiquement car le processus d'externalisation/attribution semble rester cohérent. Le "clonage" reste néanmoins perfectible et pourrait être étendu à la voix avec des ajustements similaires sur un synthétiseur vocal.
En fournissant les annotations de l'Agent-Miroir dans le corpus *POTUS*, il est possible d'évaluer les deux stratégies proposées par Rehm and André (2008). Les annotations de l'Agent-Miroir donnent une estimation des performances pour la stratégie "clone" qui pourront ensuite être comparées aux performances des règles extraites par les modèles faits dans cette thèse dès que leur évaluation aura été complétée, ce qui est une perspective à court terme de ces travaux.

Ce qu'il faut retenir :

Contributions :

- une méthode automatique d'extraction de signaux sociaux liés aux expressions faciales et à la prosodie
- un corpus d'allocutions présidentielles annoté en attitudes composé de vidéos de Barack Obama mais aussi d'un Agent-Miroir qui le "clone"
- l'évaluation de la stratégie "clone" sur ce corpus

Chapitre 4

Un modèle de fouille de séquence : la méthodologie SMART

Sommaire

AFIN de trouver l'information temporelle, une première solution consiste à utiliser les approches de fouilles de données. Les premiers résultats présents dans la littérature montrent en effet l'intérêt de ces techniques. Il s'agit d'analyser

de grands jeux de données afin d'en extraire des motifs ou associations intéressants (Masseglia et al., 2003). Les derniers développements dans ce domaine permettent de trouver des associations temporelles précises. Le but ici est d'appliquer ces techniques sur des signaux sociaux avec deux contraintes supplémentaires : obtenir un résultat propre à un phénomène affectif et pouvoir l'utiliser dans une tâche de synthèse avec un agent virtuel.

Ce chapitre va présenter les différents algorithmes pouvant répondre à cette question. Cette thèse propose la méthodologie *SMART* pour en adapter un à cette problématique. En utilisant les signaux sociaux comme des événements temporels symboliques, *SMART* trouve des associations temporelles propres à un état affectif. Cette méthode prend également en compte d'autres informations contextuelles pour garantir la généralisation de ses résultats. Ils permettent ensuite de contrôler un agent virtuel pour exprimer l'état affectif voulu. Des limitations apparaissent néanmoins dans son intégration dans une plate-forme d'agents, en particulier pour la multi-modalité, mais des solutions sont proposées pour des développements futurs.

4.1 L'exploration de données

4.1.1 Définitions

L'exploration de données correspond à l'ensemble des méthodes informatiques destinées à trouver des motifs dans de grandes bases de données. En français, les expressions *fouille de données*, *forage de données*, *prospection de données*, *extraction de connaissances à partir de données* ou l'anglicisme *data mining* y font référence. Ces terminologies soulignent l'idée d'extraction de motifs intéressants au milieu d'un vaste ensemble de données puis la transformation de ces derniers en structures intelligibles afin d'enrichir nos connaissances sur le domaine étudié.

Il s'agit de construire des modèles pour trouver des informations intéressantes à partir de critères fixés au préalable via l'analyse d'une grande quantité de données. Il est ainsi possible de trouver des groupes de données similaires (*partitionnement de données* ou *cluster analysis*), des données atypiques (*détection d'anomalies* ou *anomaly detection*) ou encore des dépendances entre elles (sous forme de *motifs séquentiels* ou *sequential pattern*, ou de *règles d'associations* ou *Association rule learning*).

Cette thèse s'est intéressée à la dynamique des liens entre des signaux sociaux multimodaux afin de pouvoir exprimer des états affectifs choisis. C'est pourquoi ces travaux se sont orientés vers la recherche de motifs séquentiels et de règles d'associations qui répondent à ces besoins. Ce domaine va être plus détaillé dans la partie 4.1.2. Les autres

branches de la fouille de données sont intéressantes pour repérer des informations générales sur un type d'interaction (f_0 moyenne par exemple) ou détecter des ruptures typiques d'une modification de l'affect mais semblent moins appropriées pour la tâche de synthèse voulue ici.

4.1.2 La recherche de règles d'associations

La recherche de *motifs séquentiels* ou de *règles d'associations* utilise l'outil statistique pour trouver des dépendances pertinentes dans les données. Elles seront représentées sous la forme de séquences ou de règles d'associations entre les différents éléments. Pour cela, des mesures d'intérêt permettent de relier des éléments ensemble sous la forme d'une implication logique "conditions $\rightarrow$ conséquence". L'exemple historique est "les hommes qui, entre 17h et 19h, achètent des couches vont acheter des bières" a ainsi modifié le placement de produits dans de nombreux magasins. En particulier, l'algorithme *APRiori* (Agrawal and Srikant, 1994) permet de détecter les occurences fréquentes de ce type dans une base de données.

La définition des modèles les rend particulièrement appropriés pour l'analyse et la prédiction du comportement humain, en particulier pour des consommateurs. Ils ont eu ainsi de nombreuses applications en marketing, fouille du web, détection d'intrusion et bio-informatique. Cette représentation est de plus très efficace pour décrire des problématiques temporelles. Plusieurs algorithmes de recherche de séquences temporelles symboliques ont été proposés. Les différentes approches vont être passées en revue afin de justifier le choix fait dans cette thèse.

Les algorithmes vont utiliser des outils statistiques comme la confiance ou le support d'un motif ou d'une association pour les créer. Pour une association $A : X \rightarrow Y$, la confiance est la probabilité sachant X de vérifier l'association A. Le support est la probabilité pour Y de vérifier l'association A. La plupart du temps, ces paramètres servent de seuils pour limiter les calculs du modèle.
Ainsi, l'algorithme *APriori*, proposé par Agrawal and Srikant (1994), utilise des seuils sur le support et la confiance d'éléments présents dans une base de données pour établir des règles d'associations sans réelle information temporelle.

Les *Systèmes Temporels Contraints* (*Temporal Constraint System* ou TCS) font le lien entre des événements et des conditions entre eux. Ainsi, en utilisant des opérateurs binaires comme contraintes, ils permettent de représenter sous forme de graphes les liens entre les événements. Un exemple basique est l'utilisation des opérateurs *avant, pendant, après*. Le TCS va se diviser en deux parties : un corps (body), composé d'événements observés, et une tête (head) dont la prédiction sera évaluée selon l'apparition du corps. Pour

cela, des informations comme la confiance, le support ou l'intervalle temporel peuvent être ajoutés. Ainsi, un TCS peut se transformer en règle d'association temporelle et inversement. L'un des premiers TCS a été défini par Dechter et al. (1991) qui limite les opérateurs binaires de contraintes à des intervalles de temps.

Différentes variantes des TCS vont être proposées comme les *Episodes* où les événements du corps ne peuvent apparaître qu'avant les événements de la tête. L'ajout de contraintes sur la distance séparant les événements va améliorer les résultats obtenus. C'est en particulier le cas de WinEpi et MinEpi (Mannila and Toivonen, 1996; Mannila et al., 1997), le premier modélisant les co-ocurences entre le corps et la tête dans une fenêtre de temps, le second proposant une fenêtre de temps pour le corps et une pour la tête et trouvant l'association entre les deux. L'utilisation de réseaux bayésiens dans les Temporal Node Bayesian Network (Arroyo-Figueroa and Sucar, 1999) va permettre de modéliser les probabilités de relations entre les éléments. Cependant, cette approche était limitée à des modèles de taille deux ou trois.

Partant de l'algorithme APriori(Agrawal and Srikant, 1994), Generalized Sequential Pattern (Srikant and Agrawal, 1996) va ajouter l'information d'ordonnancement dans les règles trouvées. Plusieurs analyses des données sont faites pour trouver des associations de plus en plus grandes jusqu'à ce que leurs fréquences d'apparition ne soit plus pertinentes. L'utilisation de la fréquence permet à l'algorithme FACE (Dousson and Duong, 1999) de discriminer l'information pertinente du bruit ambiant et d'ajouter de l'information temporelle pour trouver ainsi des associations d'événements appropriées dans les données.

Initialement, l'utilisateur supervisait les contraintes temporelles que l'algorithme avait à gérer en indiquant au modèle les intervalles de temps à considérer. Dans Chen et al. (2003) qui présente les *Time-Interval Sequential Patterns* (TIIPS) et Hirate and Yamana (2006) qui introduit les *Generalized Sequential Pattern with Item Intervals* (GSPII), les intervalles de temps sont initialement définis par l'utilisateur, soit via un Δ temps, soit par une série d'intervalles contigus et ordonnés (de la forme $[0, T_1], [T_1, T_2], [T_2, T_3]...[T_n, \infty]$. Ces systèmes étaient capables de trouver des associations intéressantes, mais sont très rigides sur les contraintes temporelles. En effet, les intervalles de temps sont pré-définis par l'utilisateur et les algorithmes ne permettent donc pas de les définir automatiquement. Ainsi, la relation temporelle trouvée est la plus appropriée dans un ensemble donnée, mais n'est pas forcément la plus pertinente pour définir le phénomène observé.

Les capacités de calcul grandissant, de nombreuses approches ont ensuite été développées sur les bases détaillées précédemment (Aggarwal and Han, 2014). En utilisant des méthodes comme les mixtures de gaussiennes, l'algorithme *expectation maximization* ou de la coloration de graphes, les différentes solutions vont chercher à déterminer des inter-

valles temporelles précis dans des jeux de données variés. Ces approches ont été utilisées dans le domaine de la santé (Ho et al., 2003; Sacchi et al., 2007; Batal et al., 2013), de la détection d'intrusions (Li et al., 2010), pour de la capture de mouvements (Li et al., 2009), . . . , illustrant ainsi le champs d'application de la recherche de règles d'associations avec une temporalité précise.

4.1.3 TITARL et les règles d'associations temporelles

Parmi les différentes solutions de fouille de données avec une analyse temporelle, l'approche proposée par Guillame-Bert and Crowley (2012), intitulée *Temporal Interval Tree Association Rules Learning* (TITARL), a servi de base à la solution proposée dans cette thèse. D'autres solutions existent (Naqvi et al., 2011; Liang et al., 2005) mais l'algorithme de Guillame-Bert propose des performances intéressantes et une implémentation claire.

Initialement utilisé et testé pour des applications diverses (habitats intelligents, domaine de la santé, applications bancaires. . .), l'algorithme a été retenu pour sa flexibilité et sa capacité à modéliser l'incertitude et l'inexactitude temporelle entre des événements temporels. En effet, en plus de trouver des relations entre des événements sous forme de règles (avant/après), il en ressort également des contraintes temporelles précises comme, par exemple, "*si il y a un événement D au temps t, il y aura un événement C entre t+5 et t+10*". Cette approche d'apprentissage temporel permet de représenter des relations temporelles imprécises et non déterministes.

TITARL présente aussi l'avantage de travailler sur des événements symboliques temporels ce qui permet de conserver un maximum d'informations pour la tâche de synthèse. En effet, comme cela est illustré dans la figure 4.1, les données numériques issues des instruments de mesure peuvent être travaillées pour être représentées dans une forme plus propice à l'exploration de données. L'utilisation de séquences d'événements symboliques temporelles pour TITARL s'applique bien aux signaux sociaux comme aux annotations. La symbolisation selon différents seuils de valeurs va permettre d'avoir des informations riches qui sont appropriées pour de la synthèse de comportements d'un agent conversationnel. L'utilisation d'événements plutôt que de séquences va laisser une certaine latitude pour trouver les associations entre les signaux.

Cette thèse propose donc d'appliquer TITARL à des signaux sociaux considérés comme des événements temporels symboliques. Au-delà de la recherche d'associations temporelles représentant une attitude sociale, le but est également d'être capable de l'appliquer à de la synthèse de comportements d'un ACA. Cette partie propose donc d'introduire TITARL tel que présenté par Mathieu Guilame-Bert pour réaliser des tâches de détection

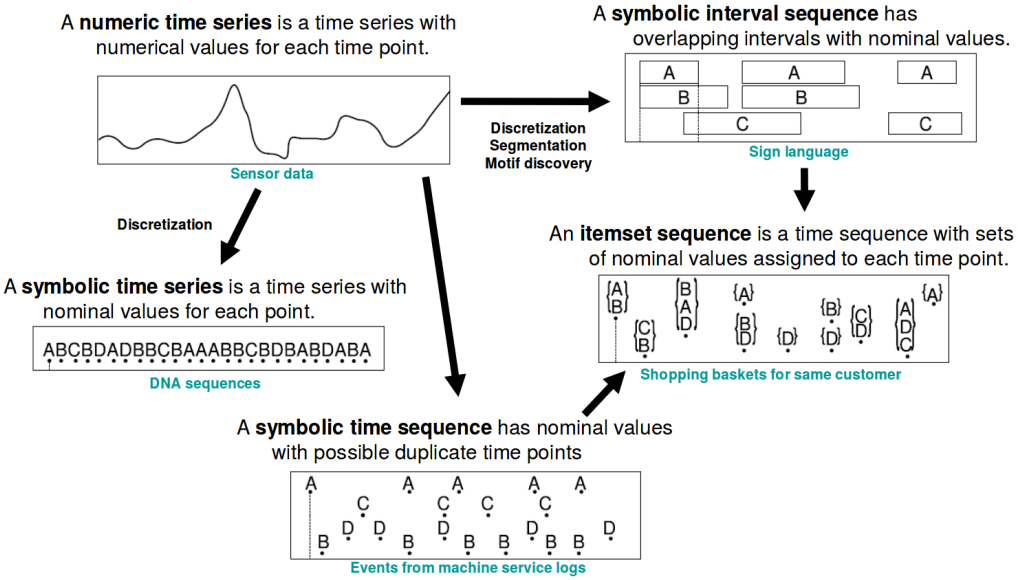

FIGURE 4.1 – Modèles communs de représentation de données temporelles pour la recherche de motifs. (issus de https://www.siam.org/meetings/sdm11/moerchen.pdf).

et de prédictions. Afin de pouvoir l'utiliser pour de la synthèse d'agents conversationnels animés, des adaptations ont été nécessaires qui sont ensuite introduites dans la partie 4.4 dédiée à son intégration dans le système SMART.

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 4.2 est : "*Une activation de l'AU 4 du locuteur à l'instant t sera suivi par une activation de l'AU 9 entre* $t+\Delta t$ *et* $t+3\Delta t$, ce qui est formalisé par la règle 4.1 :

$$AU4_{\substack{\text{activation}\\locuteur}} \xrightarrow{\Delta t, 3\Delta t} AU9_{\substack{\text{activation}\\locuteur}} \tag{4.1}$$

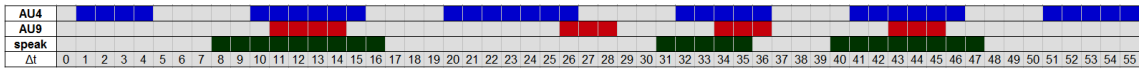

FIGURE 4.2 – Exemple de chronologie contenant les activations des AU 4 et AU 9 ainsi que les tours de parole.

Cela signifie que si le locuteur fronce les sourcils (AU 4), il plissera le nez (AU 4) 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 (4.2), notée r, traduisant que l'événement A sera suivi par l'événement B dans l'intervalle de temps $[t_{min}, t_{max}]$:

$$A \xrightarrow{\Delta t_{min}, \Delta t_{max}} B \tag{4.2}$$

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 4.3a). Son support représente le pourcentage d'événements expliqués par celle-ci (voir 4.3b). Enfin, la précision de ces règles est définie comme la dispersion de la distribution des événements A vérifiant r (voir 4.3c). TITARL assure en particulier une bonne précision de ces règles.

$$\text{confiance} = P(B(t')|A(t)), t' - t \in [\Delta t_{min}, \Delta t_{max}] \tag{4.3a}$$

$$\text{support} = \frac{\#\ B, \exists A\ tel\ que\ (A \rightarrow B)\ vrai}{\#\ B} \tag{4.3b}$$

$$\text{précision} = \frac{1}{\text{std([t'-t, } \exists \text{ A, B, } (A(t) \rightarrow B(t'))\text{ vrai])}} \tag{4.3c}$$

Le fonctionnement de TITARL est présenté dans l'image 4.3 et est divisé en deux grandes parties : un calcul brut de règles simples ou complexes suivi d'une étape d'affinage pour améliorer la pertinence des règles trouvées.
L'entrée est un ensemble d'événements temporels et une première étape va calculer des "règles simples" avec une structure comme dans l'équation 4.2 qui auront une très grande distribution temporelle. Ces règles ont donc une très grande confiance et un support important, mais une précision très mauvaise ce qui les rend non pertinentes. Par exemple, dans les habitats intelligents, si quelqu'un rentre dans une pièce, il en sera sorti après un temps infini. Cette règle est toujours vraie mais inintéressante, car elle n'apporte pas d'information. Ce problème est cependant résolu avec les deux traitements appliqués par la suite : division et raffinement. Ils visent à améliorer la précision des règles en diminuant leurs intervalles temporels tout en conservant des précisions et supports satisfaisants.

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

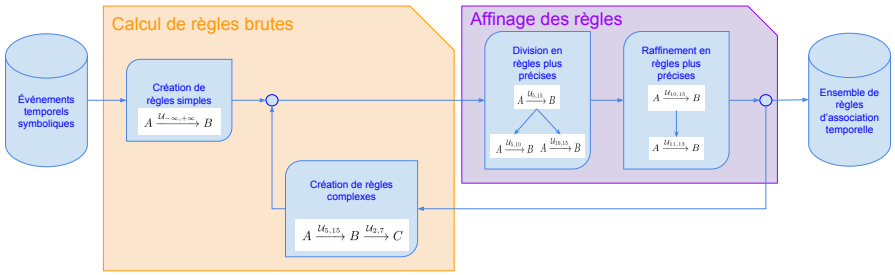

FIGURE 4.3 – Schéma de fonctionnement de TITARL

calculée et sert à la création d'un graphe en l'utilisant comme une matrice d'adjacence. Un algorithme de coloration de graphe permet de déterminer ensuite les ensembles cohérents et d'appliquer les divisions correspondantes.

L'étape de raffinement a pour but d'augmenter la précision d'une règle. Il s'agit d'observer la distribution des événements vérifiant la règle et d'appliquer un seuillage sur son histogramme de distribution. Cela permet de réduire sa variance en éliminant les valeurs limites peu présentes.

Les règles pourront ensuite être complexifiées en analysant d'autres événements à placer en tête et être stockées pour une utilisation future en reconnaissance ou en prédiction. Ainsi, un exemple d'utilisation de TITARL est d'utiliser des données physiologiques d'un patient pour anticiper un arrêt cardiaque éventuel. Plus d'informations sur le fonctionnement de TITARL peuvent être trouvées dans le papier de Guillame-Bert and Crowley (2012).

4.2 Le système SMART

TITARL est donc un bon outil pour trouver des associations temporelles entre des événements afin de pouvoir faire de la reconnaissance ou de la prédiction. La problématique de cette thèse est de pouvoir synthétiser des attitudes sociales différentes à partir de signaux sociaux. En repartant de la structure de TITARL, des modifications ont été apportées pour créer le système *SMART* pour *Social Multimodal Association Rules with Timing*. Capable de trouver des règles d'associations temporelles spécifiques de différentes expressions d'attitudes à partir de données extraites automatiquement, *SMART* les transforme en information qui peut être synthétisée avec un agent virtuel comme cela est illustré dans la figure 4.4.

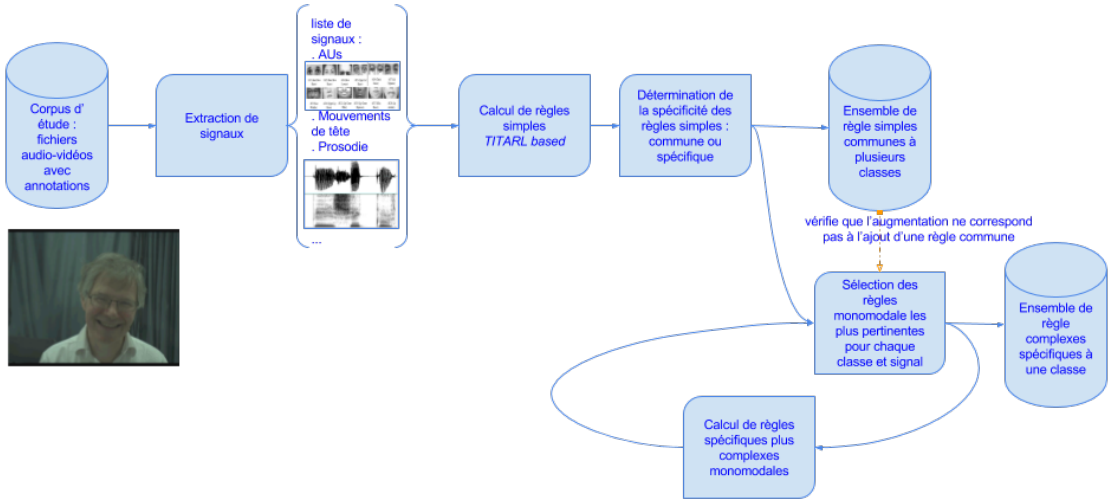

FIGURE 4.4 – Schéma de fonctionnement de SMART

Une première étape importante consiste à extraire automatiquement des données des signaux sociaux (voir 3.2) et à les transformer en événements symboliques temporels (voir 4.2.1). L'algorithme TITARL est ensuite utilisé à deux reprises pour le calcul des règles simples et des règles complexes (i.e. des règles comprenant plus d'associations). Lors de ces étapes, des stratégies de calcul ont été choisies afin de gérer la multimodalité des signaux (voir 4.2.2) et d'assurer la pertinence des règles trouvées pour exprimer différents états affectifs (voir 4.2.3). Enfin, un travail est effectué afin de transformer les règles d'associations temporelles trouvées pour la synthèse avec un agent (voir 4.2.4 et 4.2.5). Ces différentes modifications vont maintenant être passées en revue et détaillées.

4.2.1 Symbolisation des signaux extraits automatiquement

Pour pouvoir utiliser l'algorithme de fouille de données sur les signaux sociaux, ces derniers doivent être transformés en événements temporels symboliques. Afin d'être indépendants du protagoniste, les signaux sont centrés-normalisés, puis, le procédé de symbolisation est appliqué.

La symbolisation des signaux s'est faite simplement en fonction de leur pourcentage de variation par rapport à la moyenne des valeurs observées, voir eq.(4.4) pour les signaux

continus (f_0, angle de la tête).

$$\text{symbole(valeur)} = 100 * (\frac{\text{valeur}}{\text{moyenne des valeurs}} - 1) \tag{4.4}$$

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

Enfin, deux stratégies ont été testées sur la définition des événements symboliques. La première solution consiste à créer un événement à chaque mesure du signal d'entrée, ce qui en pratique correspond à la vitesse de trame. La seconde consiste à ne regarder que les transitions entre les différents symboles d'un signal (tête de 0% à 10% par exemple). Cette deuxième solution s'avère satisfaisante pour trouver des associations efficaces tout en garantissant un temps de calcul réduit. Cette symbolisation sera donc retenue dans la plupart des applications menées.

4.2.2 Stratégie de calcul des règles d'associations temporelles et multimodalité

TITARL est modifié et adapté pour qu'il corresponde mieux à la problématique de cette thèse en travaillant sur les signaux pris en compte pour le calcul des règles.

TITARL est initialement capable de trouver, en utilisant différents signaux, des règles d'associations temporelles caractéristiques dans un jeu de données. Dans le cadre des signaux sociaux, de nombreuses modalités sont disponibles. Il peut s'agir de plusieurs *action units*, de mouvements de tête, d'informations prosodiques comme la f_0 ou la prise de parole. En utilisant directement TITARL, les règles trouvées peuvent donc mélanger ces signaux.

Les premiers tests effectués ont soulevé deux problèmes liés à cette approche. La structure de TITARL fait qu'en utilisant des règles comprenant de nombreux signaux, le temps de calcul des associations augmente significativement. Par ailleurs, le but de SMART est de proposer des règles utilisables pour générer le comportement d'un ACA. Des règles multimodales se sont avérées compliquées à transposer pour être synthétisées dans le comportement de l'agent. La synchronisation des signaux de manière précise avec l'architecture *SAIBA* [1] (*Situation, Agent, Intention, Behavior, Animation*) s'est avérée complexe, en particulier à cause des différentes échelles de temps. *SAIBA* est l'architecture de référence dans le domaine de la génération d'agents virtuels et est basée sur différents langages de type *XML*, en particulier le *BML* pour le comportement et le *SSML* pour la voix qui seront

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

détaillés dans la partie 4.3.
Par exemple, les expressions faciales sont mesurées en secondes mais le contour prosodique évolue selon le pourcentage temporel lors de la prononciation de la phrase (voir figure 4.7). Ces différentes mesures du temps lors du contrôle des modalités de l'agent ont empêché l'utilisation directe de SMART. Des développements supplémentaires sont donc à prévoir pour satisfaire cette architecture.

Pour palier ce problème, les règles ont donc été calculées pour une seule modalité à la fois. Cela permet également d'assurer que de l'information est trouvée pour chaque signal présenté. Ensuite, à la fin de chaque étape de calcul, les co-occurrences de règles sont évaluées, et ce, pour toutes les modalités. Ainsi, il est possible de retrouver des informations multimodales et d'effectuer une synthèse sur l'agent en utilisant divers signaux.

4.2.3 Sélection des règles pertinentes

Le score indiqué dans l'équation 4.5 a été défini par Guillame-Bert and Crowley (2012) comme une pondération entre différentes caractéristiques d'une règle r dont sa confiance, son support et son intervalle temporel, défini par ses bornes t_{min} et t_{max}. Il permet ainsi de classer les règles calculées en fonction de leurs pertinences.

$$Score = \frac{conf_r^4 . supp_r^2}{t_{max} - t_{min}} \tag{4.5}$$

Une limite de ce score est qu'il n'est pas discriminant entre différents états, qui sont ici des attitudes mais qui pourrait correspondre à d'autres tâches. Pour permettre cette différence, les données en entrée sont considérées comme appartenant à S différentes sessions $s_0, s_1, s_2, ..., s_S$, où différentes attitudes apparaissent $att_0, att_1, att_2, ..., att_N$. Une session correspond à un tour de parole d'un locuteur lors d'une interaction face à face. Cela permet de comparer l'information prosodique et l'information visuelle. Elle est ainsi délimitée par la prise de parole et la pause à la fin. Pour une règle représentant l'attitude att_i, le score présenté dans 4.5 est calculé pour toutes les sessions où l'attitude a_i est exprimée, noté sc_{att_i}. En s'intéressant au score sur les sessions et non l'ensemble des données, la définition même de la confiance et du support en fait une représentation de la fréquence d'apparition de l'association pour une attitude. La règle est ensuite évaluée pour les autres attitudes et les scores correspondant sont calculés. Ainsi, la règle est considérée comme pertinente pour l'attitude i si elle vérifie l'équation 4.6 où le seuil sur le score à été fixé au préalable (1.2).

$$\forall j \neq i, sc_{att_i} > \text{seuil score} * sc_{att_j} \tag{4.6}$$

Ainsi, des règles communes aux attitudes ne seront pas retenues. Elles peuvent par exemple correspondre aux mouvements de la mâchoire lors de la production de la parole.

Par ailleurs, si les données en entrée contiennent différents protagonistes, notés pr_0, pr_1, ..., pr_m, un coefficient de variation est utilisé pour assurer la généralisation des résultats obtenus. Ce coefficient, présenté dans 4.7, permet ainsi de garantir que la règle apparaît chez plusieurs personnes : une valeur élevée indiquant que la règle est spécifique à quelqu'un là où un coefficient proche de zéro assure la généralisation. Comme le but est d'obtenir des règles pour de la synthèse, les associations les plus générales sont retenues.

$$\text{var coef} = \frac{\sum_{i=1}^{m} (conf_r^4.supp_r^2 - conf_{r,pr_i}^4.supp_{r,pr_i}^2)^2}{S.conf_r^4.supp_r^2} \tag{4.7}$$

A chaque étape de calcul, SMART vérifie donc que la règle est bien significative pour une attitude tout en restant généralisable chez plusieurs personnes, assurant ainsi sa pertinence pour la synthèse de l'agent.

Enfin, lors du calcul de règles plus compliquées en taille, une vérification assure qu'il ne s'agisse pas de la concaténation d'une règle spécifique avec une règle commune à plusieurs attitudes. Pour cela, lors du calcul des règles simples, l'ensemble des règles communes est stocké en mémoire. A la fin de chaque étape de calcul de règles complexes, la base de règles simples communes est interrogée afin d'assurer la pertinence de la complexification calculée.

4.2.4 Consistance des règles

La seconde adaptation proposée s'attaque au problème de cohérence des règles obtenues pour l'étape de génération. 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'agent conversationnel animé. Cela peut être expliqué par la structure de TITARL pour le calcul des règles et par les signaux étudiés.

En effet, nos signaux ne sont pas à valeurs binaires ce qui augmente le risque de calcul de règles inconsistantes. 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 (4.8).

$$\mathrm{S}_{v_1 \text{ à } v_2} \xrightarrow{\Delta t_{min};\Delta t_{max}} \mathrm{S}_{v_1 \text{ à } v_2} \tag{4.8}$$

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 va entraîner une diminution du support et de la confiance de la règle, diminuant également le score au point que *SMART* ne retiendra pas 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 sur les valeurs de l'AU_1 (haussement de sourcils), un résultat possible est : *un événement AU_1 désactivée à une faible activation est suivi 3 secondes plus tard par un événement AU_1 désactivée à une faible activation*. Pour faire de la synthèse à partir de cette règle, le système doit intégrer des renseignements sur le moment de la désactivation de l'AU_1. L'algorithme de détection des *action units* peut avoir trouvé des activations fortes et la continuité dans la transition ne peut pas être trouvée par TITARL.

Pour corriger ce problème, une analyse des transitions possibles permet de retenir les candidats pertinents et de forcer ensuite leur ajout dans l'arbre d'associations des règles. La règle de l'équation 4.8 devient alors celle présentée dans l'équation 4.9.

$$\mathrm{S}_{v_1 \text{ à } v_2} \xrightarrow{\Delta t_{min_1};\Delta t_{max_1}} \mathrm{S}_{v_2 \text{ à } v_1} \xrightarrow{\Delta t_{min_2};\Delta t_{max_2}} \mathrm{S}_{v_1 \text{ à } v_2} \tag{4.9}$$

La cohérence de la règle est ainsi assurée pour l'étape de génération car les événements la composant ont des changements d'états continus et cohérents.

4.2.5 Application à la synthèse

La dernière étape de notre système SMART consiste à transformer la règle d'association temporelle en fichiers *BML* et *SSML* pour la synthèse. Le Behavior Markup Language (*BML*) est un langage de type XML qui permet le contrôle du comportement verbal et non-verbal d'un agent. 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.

Ces fichiers indiquent la temporalité de différents signaux sociaux exprimés par l'agent virtuel 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 4.5), trouvant ainsi simplement les temps de transitions nécessaires au *BML* et *SSML*. Dans de futures versions, les distributions des occurrences pourront être utilisées pour sélectionner différents temps de transitions et introduire ainsi plus de variabilité dans la génération des animations. Cela est d'autant plus

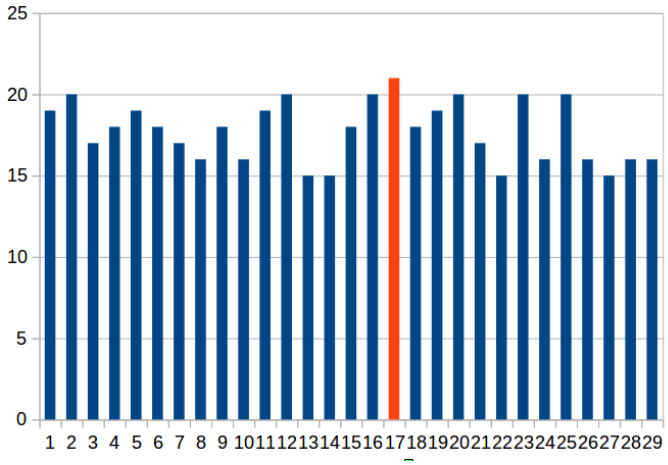

FIGURE 4.5 – Exemple de distribution des occurrences des événements vérifiant une règle. Le Δ_t ayant le plus d'occurrences est affiché en orange.

vrai que d'autres temps de transitions non retenus peuvent avoir un nombre d'occurrences très proche du maximum.

4.3 Validations : études selon différents signaux sociaux et différentes échelles de temps

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. Les résultats sont ensuite synthétisés en respectant les formats *BML/SSML* utilisés dans la plate-forme d'agents. La première se concentre sur des ensembles d'*action units* tandis que la seconde combine *action units* et événements prosodiques.

4.3.1 Étude 1 : *action units*, mouvements de tête et secondes

Cette première étude met l'accent sur les *action units* correspondant au sourire (AU_6, AU_{12}) et aux mouvements de sourcils (AU_{1+2}, AU_4) afin de tester *SMART* sur ces signaux sociaux spécifiques. Cela permet de valider la méthode en comparant ses résultats aux liens trouvés dans Ochs and Pelachaud (2012); Ravenet et al. (2013) sur des études d'agents conversationnels animés. Ces articles soulignent qu'une attitude amicale comporte de nombreux sourires alors qu'une attitude hostile est exprimée par de nombreux froncements de sourcils.

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

TABLEAU 4.1 – Exemples de règles trouvées par TITARL. La première partie montre les liens trouvés entre les sourires (AU_6 et AU_{12}) 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, leur confiance (colonne c), leur support (su), leur score (sc)

Le tableau 4.1 montre des règles avec leurs confiances, supports, scores et ratios de fréquence. Il s'agit de règles avec l'un des meilleurs scores et un ratio de score spécifique 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 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.

Ainsi, ces résultats sont en accord avec la littérature et ajoutent à celle-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 précise la durée de ces signaux sociaux qui est une information très importante pour la génération d'une attitude pour un agent virtuel.

Les meilleures règles trouvées par le système permettent ensuite de générer le comportement d'agents conversationnels animés, dont des exemples sont visibles ici [2]. Comme présenté précédemment, les *BML* ont été générés automatiquement à partir des résultats trouvés afin d'être utilisés en entrée de la plate-forme Greta (Pecune et al., 2014). Le but est d'évaluer la perception de l'attitude de l'agent. Comme le but ici est de faire de la synthèse, les règles ont été calculées en se limitant aux *action units* contrôlables dans la plate-forme. Par exemple, l'AU_9 (plissement du nez) n'est pas implémentée dans la plate-forme et n'a donc pas été étudiée.

2. https://youtu.be/O2EPivej99Y

	Min.	1st Qu.	Mediane	Moyenne	3rd Qu.	Max
Poppy	2.000	3.000	3.000	3.367	4.000	5.000
Spike	1.000	2.000	3.000	2.5	3.000	5.000

TABLEAU 4.2 – Résumé des évaluations sur les vidéos générées à partir de Poppy (amical) ou Spike (hostile).

L'étude a été menée comme suit : les règles d'associations temporelles ont été apprises sur la base *Semaine-SAL* en mode auditeur pour Poppy (amical) et Spike (hostile). Les calculs ont été arrêtés après 3 itérations d'addition et les 3 règles aux scores les plus importants ont été retenues. Ces 6 règles comportent donc des séquences d'*action units* et de mouvements de tête avec une information temporelle précise. Les occurences des signaux étudiés permettent la transposition en *BML* qui ont ensuite été utilisés pour générer les vidéos correspondantes. L'évaluation perceptive des vidéos obtenues a été réalisée à l'aide de la plate-forme *crowdflower*[3]. 97 jugements de 62 participants ont ainsi été recueillis : il leur a été demandé d'évaluer leur perception de l'hostilité ou de l'amicalité de l'agent ainsi que leur confiance en leur jugement grâce à des échelles de Likert de taille 5. Par exemple, 1 signifiait que l'agent paraissait très hostile et 5 très amical.

Un résumé général des jugements des vidéos est visible dans le tableau 4.2. Pour les règles obtenues à partir de Spike (hostile), le jugement moyen est de 2,5 mais surtout le 3ême quartile est à 3 ce qui signifie que 75% des jugements étaient entre 1 et 3. De même, pour Poppy (amical), la moyenne est de 3.367 et le 1 quartile à 3. Ces résultats indiquent donc une tendance générale selon laquelle les règles synthétisées à partir des vidéos de Poppy ont bien été perçues comme amicales et celles de Spike comme hostiles.

Une analyse statistique plus poussée a été menée en suivant les recommandations de Motulsky (2013). Comme un test de Shapiro-Wilk indiquait que les jugements ne suivaient pas une distribution normale (p-value$< 10^{-16}$), un test de Mann-Withney's U a évalué la différence dans les réponses et a indiqué un effet de groupe significatif (p$= 9.10^{-5}$). Cela confirme que les vidéos dont la synthèse est inspirée de Poppy ont obtenu un score plus élevé que celles issues de Spike. Ces résultats sont illustrés dans le graphique de la figure 4.6 qui valide que les vidéos construites à partir de règles d'associations temporelles issues de Poppy sont perçues comme plus amicales que celles de Spike, et ce avec un système de synthèse très sommaire.

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

4.3.2 Étude 2 : contours prosodiques, fréquence fondamentale et pourcentage

Cette étude explore une autre application de SMART dans le but de colorer la voix d'un agent conversationnel animé en fonction de l'attitude planifiée. La synthèse vocale d'un agent, en particulier le contrôle de sa prosodie, peut se faire avec la norme *SSML*[4]. Les contours prosodiques sont donc définis selon cette norme 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 4.7). Dans cette application, ce texte correspond à une phrase. Le second, *y*, est une valeur cible à atteindre pour la F_0. Ici, la valeur cible, *y*, correspond à 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 4.7 qui présente le paramétrage d'une phrase. 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 %.

Les descripteurs prosodiques décrits dans 3.2 ont été extraits des enregistrements des opérateurs de la base de données SAL-SOLID SEMAINE. Ces descripteurs sont ensuite transformés en événements symboliques temporels 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. Ces événements temporels symboliques sont analysés par SMART pour

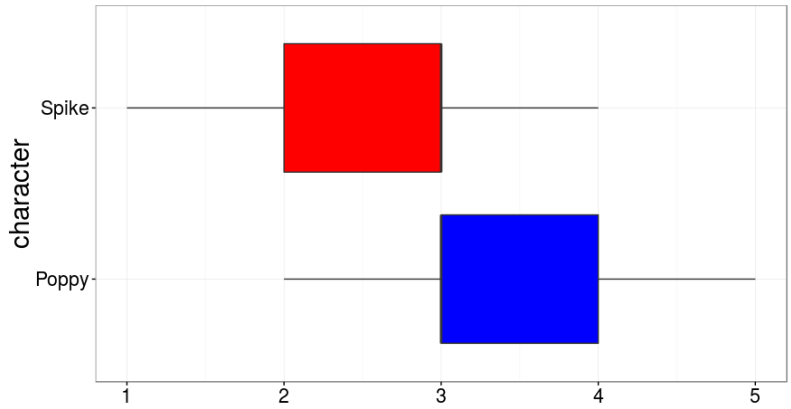

FIGURE 4.6 – Résultats de l'évaluation perceptive des différentes vidéos générées à partir des règles apprises sur Spike (hostile) et Poppy (amical).

4. https://www.w3.org/TR/speech-synthesis/

	rule ($body \xrightarrow{\Delta t_{min};\Delta t_{max}} head$)	contour
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	"(15%,50%)(73%,0%)(80%,-10%)(90%,0%)"
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	"(15%,-20%)(81%,-30%)(88%,-30%)(96%,-20%)"
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	"(60%, -20%)(69%, -20%)(74%, -20%)(87%, -10%)"
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	"(61%, 10%)(75%, 10%)(80%, 10%)(93%, 20%)"
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	"(52%, -40%)(75%, 10%)(90%, -40%)(95%, -30%)"
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	"(22%, -40%)(60%, -30%)(75%, -30%)(79%, -30%)"
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	"(3%, -20%)(12%, -10%)(90%, -10%)(95%, -10%)"
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	"(51%, -20%)(64%, -20%)(86%, -20%)(93%, -20%)"

TABLEAU 4.3 – Exemples de contours prosodiques de taille 4 trouvés 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

chercher 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 les résultats, une validation-croisée a évalué les performances des règles sélectionnées dans une tâche de reconnaissance. Le but était d'assurer que les règles sélectionnées permettent une discrimination correcte et de trouver le seuil optimal à appliquer au ratio de fréquence. Une reconnaissance correcte est obtenue avec un taux de validation de 75% pour un seuil de fréquence à 0.8. Des contours prosodiques au format de la norme *SSML* sont ainsi extraits dont des exemples sont présentés dans le tableau 4.3.

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

```
<?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 4.7 – Exemple de contour prosodique défini avec la norme *SSML*

dans la figure 4.8. 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, les valeurs des éléments composant les règles liées à Poppy sont généralement plus importantes que chez Spike. Cela est visible dans le tableau 4.3 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 cette étude, une évaluation perceptive a été menée en générant grâce à un synthétiseur vocal des phrases prononcées avec les contours prosodiques correspondant aux meilleures règles. Pour cela, deux phrases affirmatives et deux phrases interrogatives ont été sélectionnées dans les transcriptions originales de SEMAINE-DB, une prononcée par Spike et une par Poppy. Mary TTS [5] (Modular Architecture for Research on speech Synthesis) a servi 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 cachés (MMC). La synthèse a été faite avec 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.

Cette évaluation prend l'ensemble de phrases à évaluer et, pour chacune d'entre elles, les synthétise avec les paramètres par défaut du synthétiseur afin d'obtenir des fichiers dits

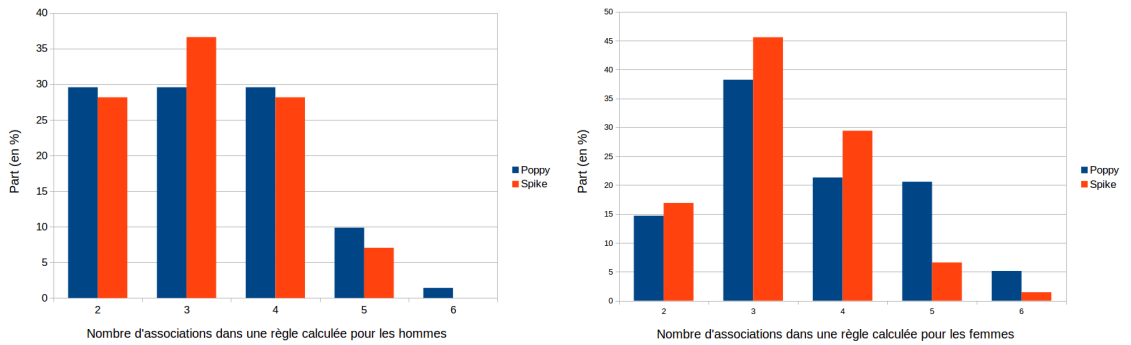

FIGURE 4.8 – Nombre d'associations trouvées dans les règles selon le personnage joué

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

"neutres".

Puis les fichiers audio avec la F_0 moyenne calculée pour Poppy ou pour Spike, sans toucher aux contours prosodiques, sont générés ce qui donne deux références, nommées *F0Spike* et *F0Poppy*.

Enfin, la synthèse prend en compte aussi les contours prosodiques calculés par SMART pour générer *ContourSpike* et *ContourPoppy*. Cela donne un ensemble de fichiers audio qui sont ensuite évalués en attitude sociale et en réalisme via la plate-forme Crowdflower. Ici, seule la synthèse vocale est analysée : les fichiers audio sont jugés sans utilisation d'agents virtuels. Les annotateurs devaient résider dans des pays anglophones. Via des échelles de Likert en 7 points, ils ont noté en amicalité et en réalisme les synthèses vocales de deux phrases, une affirmation et une interrogation, prononcées avec une voix féminine ou une voix masculine. 283 jugements ont ainsi été obtenus, visibles dans la figure 4.9.

Leur analyse montre tout d'abord que les phrases "brutes", sans modifications, ont été perçues comme légèrement amicales (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 a avoir un rendu plus hostile. Cependant, dans les deux cas, l'effet n'est pas significatif.

Une première piste pour comprendre cette absence d'effet est suggérée par Bawden et al. (2015) et consiste à regarder l'acte de dialogue. En effet, les tendances sont plus cohérentes avec ce qui est attendu 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 vues plus amicales que pour le neutre. Les résultats ont également 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 [6] ont été notés 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 comme amicaux. A

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

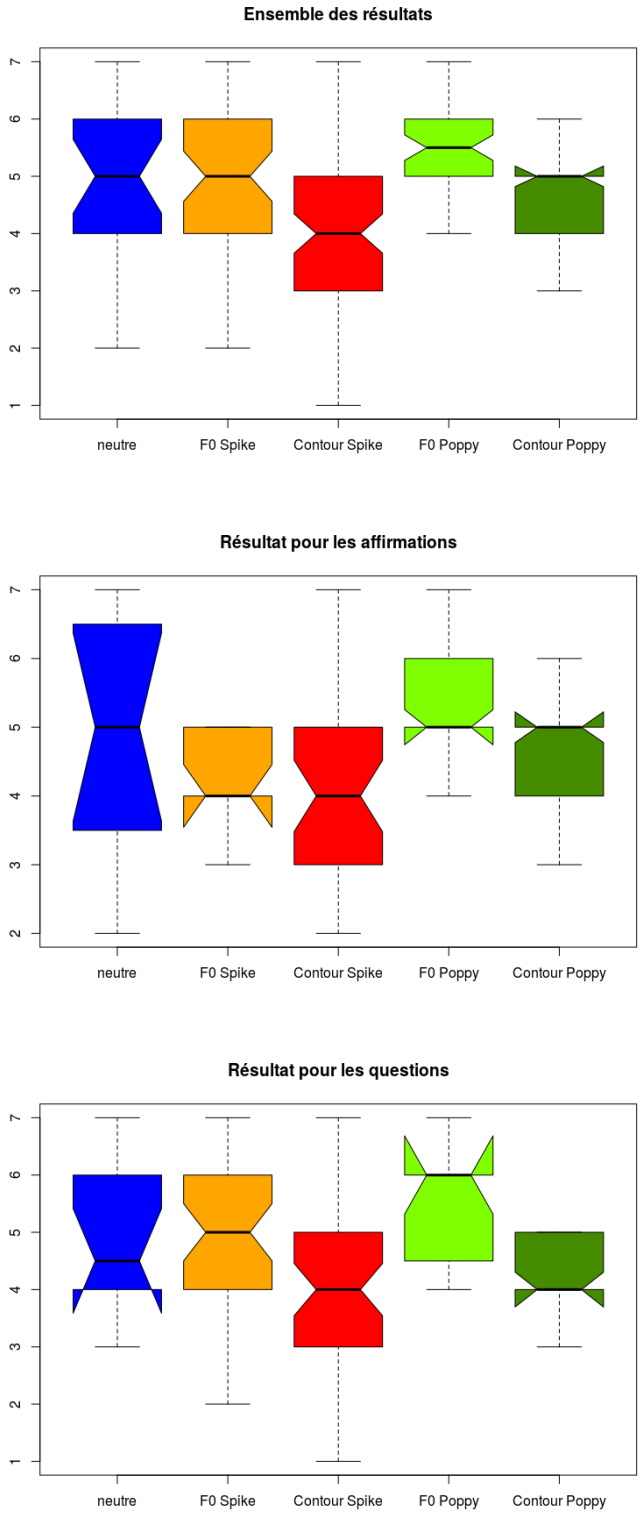

FIGURE 4.9 – Graphiques représentant l'évaluation des fichiers en amicalité, de 1 très hostile à 7 très amical. En bleu, synthèse sans modification ; En orange, synthèse avec la F_0 de Spike ; En rouge, synthèse avec les contours de Spike ; En vert clair, synthèse avec la F_0 de Poppy ; En vert foncé, synthèse avec les contours de Poppy

contrario, ces synthèses [7], basées également sur des contours appris sur Poppy, ont été jugées peu réalistes et hostiles. La voix est très mécanique et rend donc l'écoute désagréable.

Afin d'éviter cette limitation, une seconde stratégie a été employée : évaluer des fichiers audio de voix dont les contours prosodiques ont été modifiés précisément grâce à Promo [8]. 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 and Weenink, 2017). L'impact des contours prosodiques dans la perception de l'attitude sociale peut ainsi être évalué précisément. De la même façon, 120 utilisateurs anglophones de la plate-forme Crowdflower ont estimé leur perception de l'attitude sociale exprimée dans chaque fichier audio. Chaque sujet en avait 4 à évaluer : un extrait audio avec un contour plat et une F_0 amicale (i.e. la F_0 moyenne calculée sur les sessions Poppy), un extrait audio avec un contour plat hostile (i.e. avec la F_0 moyenne calculée sur les sessions Spike), un extrait audio avec un contour prosodique amical (avec F_0 moyenne et contour prosodique calculé sur Poppy) et un extrait audio avec un contour prosodique hostile (avec F_0 moyenne et contour prosodique calculé sur Spike) dont les résultats sont visibles dans la figure 4.10.

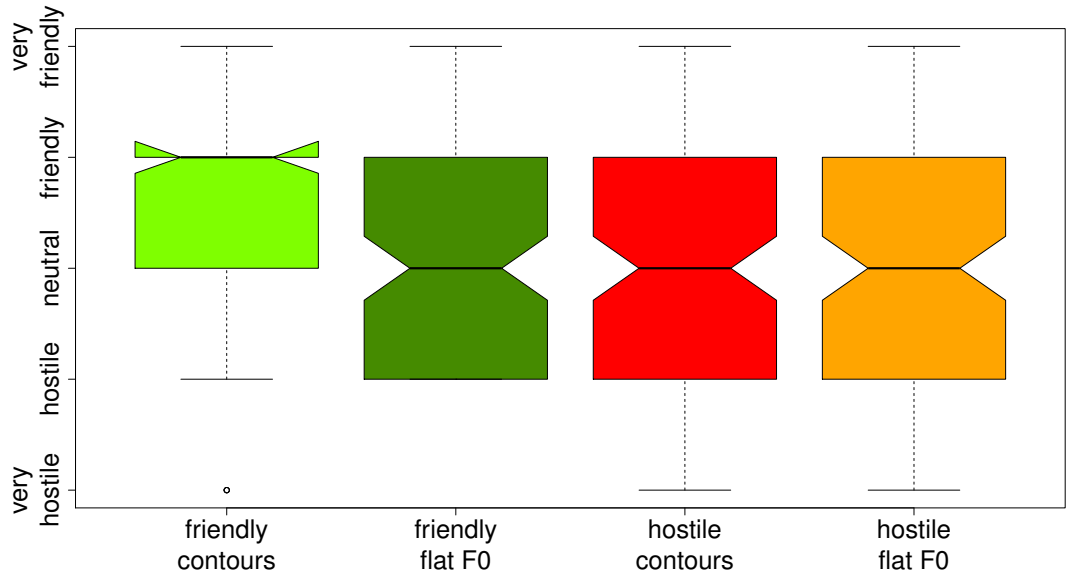

FIGURE 4.10 – Deuxième évaluation : résultats d'étude perceptive sur les audio aux contours prosodique contrôlés précisément. En vert clair, le contour amical, en vert foncé, juste la F_0 amicale, en rouge, le contour hostile et en orange, la F_0 hostile

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

8. https://github.com/timmahrt/ProMo

Toujours en suivant les recommandations de Motulsky (2013), un test de Shapiro-Wilk sur ces résultats a indiqué une distribution non normale des réponses donc un test de Man-Withney'U a permis de comparer la médiane des perceptions de chaque cas. En ce qui concerne la perception des contours prosodiques amicaux, un effet significatif est observé 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). Une analyse plus détaillée pour chaque sujet sur l'évolution de sa perception entre un contour plat et un contour prosodique montre 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 a suscité plus de variations dans les jugements des observateurs. Ces résultats restent en accord avec la littérature qui indique que la variation de F_0 peut influencer la perception de l'amicalité mais a un effet moindre sur l'hostilité.

4.4 Multimodalité : limitations et solutions

L'application de SMART pour la synthèse a été explorée afin de combiner les différents signaux disponibles et faire ainsi une synthèse conjointe de la voix, des expressions faciales et des mouvements de tête. Illustrée dans la figure 4.11, deux stratégies ont été explorées qui correspondent aux approches proposées par Snoek et al. (2005) pour la fusion de modalité pour de l'analyse. Ils définissent ainsi :

fusion précoce : Schéma de fusion qui intègre des fonctionnalités unimodales avant d'apprendre des concepts (*Early fusion* en anglais).

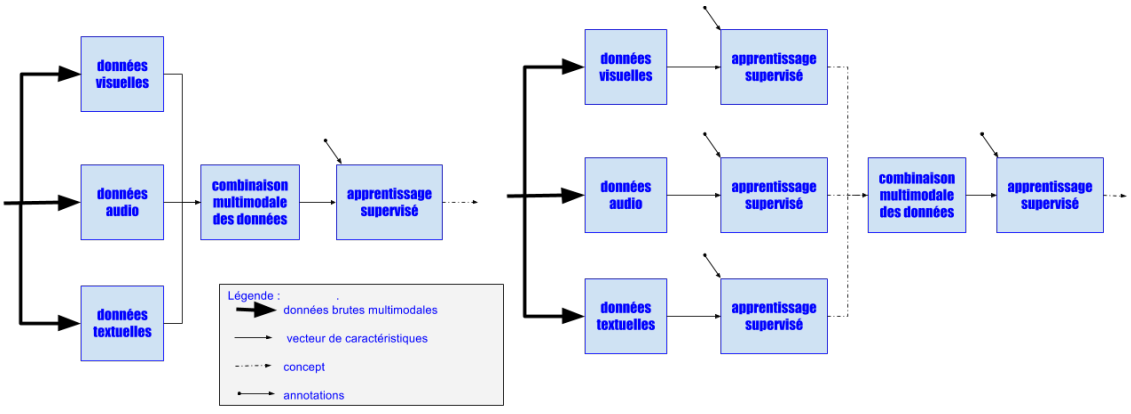

FIGURE 4.11 – Illustrations des deux stratégies de fusion de modalités, *précoce* à gauche, *tardive* à droite. Image inspiré de Snoek et al. (2005)

fusion tardive : Schéma de fusion qui réduit d'abord les caractéristiques unimodales à des scores de concepts appris séparément, puis ces scores sont intégrés pour apprendre des concepts (*Late fusion* en anglais).

Ces deux approches ont été explorées avec *SMART* et les résultats présentés ici portent sur l'analyse du corpus *POTUS*. Le changement de corpus d'étude est motivé par la qualité des signaux extraits et la précision des annotations recueillies. Cela permet également d'illustrer la généralisation de *SMART*.
Les annotations en amicalité ont été divisées en deux classes de tailles égales, *faible amicalité* et *forte amicalité*, et SMART a analysé les signaux sociaux correspondants. Une première approche a consisté à utiliser les mêmes paramètres que pour la recherche de contours prosodiques (cf. section 4.3.2 : les règles débutent par une prise de parole et se terminent avec une pause dans la voix, l'échelle de temps est en pourcentage d'avancée dans l'énoncé). L'idée était d'utiliser les actes de dialogues comme pivot pour la multimodalité.

Pour la *fusion précoce*, les règles d'associations temporelles vont pouvoir comporter des signaux de toutes les modalités. Les meilleurs résultats, spécifiques à une classe, sont visibles dans le tableau 4.4. Ces résultats soulignent deux difficultés par rapport à la multimodalité. Tout d'abord, au niveau de la cohérence des règles pour la synthèse, des ajustements doivent être faits au niveau de la fermeture des règles pour trouver toutes les transitions. Le tableau 4.4 présente un ensemble de résultats de contours multimodaux trouvés de taille 6. Il est intéressant de constater que pour la tâche relativement subtile de différencier une faible d'une forte amicalité, un nombre conséquent de règles a été trouvé (Plus de 700 mais avec des redondances selon la taille : les petites règles servant de base aux grandes). Cependant, il apparaît que les règles ne comportent que très peu de signaux différents : elles sont généralement mono-modales ou ne comportent que 2 modalités et une seule occurrence de l'une d'entre elles. Ce nombre limité d'associations trouvées dans le cas de la *fusion précoce* rend ces résultats assez peu intéressants pour une tâche de synthèse car, dans la plupart des cas, ils consistent en un contour prosodique et une seule modification d'expression faciale.

	rule ($body \xrightarrow{\Delta t_{min};\Delta t_{max}} head$)	**score**
faible amicalité	début parole $\xrightarrow{49\%;52\%} f_0(-20\%) \xrightarrow{2\%;8\%} f_0(-20\%) \xrightarrow{2\%;11\%} f_0(-20\%) \xrightarrow{2\%;5\%} f_0(-20\%) \xrightarrow{0\%;6\%} f_0(-20\%) \xrightarrow{3\%;12\%} f_0(-30\%) \xrightarrow{21\%;30\%}$ fin parole	$6,2.10^{-12}$
faible amicalité	début parole $\xrightarrow{15\%;18\%} f_0(+0\%) \xrightarrow{1\%;10\%} f_0(+0\%) \xrightarrow{28\%;31\%} f_0(-10\%) \xrightarrow{0\%;7\%} f_0(+0\%) \xrightarrow{8\%;17\%} f_0(+0\%) \xrightarrow{0\%;9\%} AU4_{0\text{ to }1} \xrightarrow{21\%;30\%}$ fin parole	$1,0.10^{-12}$
faible amicalité	début parole $\xrightarrow{15\%;18\%} f_0(+0\%) \xrightarrow{1\%;10\%} f_0(+0\%) \xrightarrow{28\%;31\%} f_0(-10\%) \xrightarrow{0\%;7\%} f_0(+0\%) \xrightarrow{8\%;17\%} f_0(+0\%) \xrightarrow{2\%;6\%} rot(\text{tête})_{Y:10\text{ to }20} \xrightarrow{23\%;31\%}$ fin parole	$3,8.10^{-13}$
forte amicalité	début parole $\xrightarrow{0\%;5\%} AU4_{1\text{ to }0} \xrightarrow{12\%;21\%} f_0(+0\%) \xrightarrow{1\%;8\%} f_0(+0\%) \xrightarrow{26\%;35\%} f_0(-10\%) \xrightarrow{1\%;4\%} f_0(-10\%) \xrightarrow{0\%;9\%} f_0(-10\%) \xrightarrow{41\%;47\%}$ fin parole	$1.2.10^{-12}$
forte amicalité	début parole $\xrightarrow{0\%;5\%} AU1_{0\text{ to }1} \xrightarrow{0\%;7\%} AU1_{1\text{ to }2} \xrightarrow{25\%;33\%} f_0(-10\%) \xrightarrow{37\%;46\%} f_0(-10\%) \xrightarrow{1\%;10\%} f_0(-10\%) \xrightarrow{1\%;10\%} f_0(-20\%) \xrightarrow{13\%;17\%}$ fin parole	$5,0.10^{-13}$
forte amicalité	début parole $\xrightarrow{0\%;5\%} AU4_{1\text{ to }0} \xrightarrow{12\%;21\%} f_0(+0\%) \xrightarrow{1\%;8\%} f_0(+0\%) \xrightarrow{7\%;15\%} f_0(-10\%) \xrightarrow{16\%;25\%} f_0(-10\%) \xrightarrow{0\%;8\%} AU12_{0\text{ to }1} \xrightarrow{39\%;47\%}$ fin parole	$4,7.10^{-13}$

TABLEAU 4.4 – Exemples de règles multimodales trouvées par SMART sur des tours de paroles.

Face à ces limites, des solutions en *fusion tardive* ont été explorées. La méthode actuellement en développement devrait permettre de générer efficacement des comportements d'agents exprimant des attitudes en se reposant sur différents signaux multimodaux. Elle se décompose en deux étapes : premièrement, des règles d'associations temporelles monomodales sont extraites selon la méthode classique afin de trouver l'équivalent des contours prosodiques pour chaque signal. Les co-occurences d'apparition des règles sont ensuite calculées afin de trouver l'information multi-modale.

Par exemple, pour la règle du contour prosodique R1 :

$$\text{R1 : début parole} \xrightarrow{1\%;10\%} f_0(+20\%) \xrightarrow{0\%;24\%} f_0(+40\%) \xrightarrow{5\%;31\%} f_0(+10\%) \xrightarrow{0\%;3\%} f_0(+10\%) \xrightarrow{5\%;19\%} f_0(-10\%) \xrightarrow{1\%;20\%} f_0(+20\%) \xrightarrow{32\%;61\%} \text{fin parole}$$

les règles R2 et R3 ont environ 16% de chance d'apparaître dans le même temps. Cependant, R2 et R3 n'ont que 8% de chance d'être co-occurentes. Pour des raisons de clarté, les règles ici ne sont pas présentées après les étapes de fermetures pour la cohérence.

$$\text{R2 : début parole} \xrightarrow{0\%;21\%} AU4_{0\text{ to }1} \xrightarrow{1\%;22\%} AU4_{1\text{ to }2} \xrightarrow{0\%;6\%} AU4_{2\text{ to }1} \xrightarrow{23\%;51\%} AU4_{0\text{ to }1} \xrightarrow{3\%;32\%} AU4_{1\text{ to }0} \xrightarrow{0\%;28\%} AU4_{0\text{ to }1} \xrightarrow{0\%;28\%} \text{fin parole}$$

$$\text{R3 : début parole} \xrightarrow{2\%;13\%} AU1_{1\text{ to }2} \xrightarrow{4\%;25\%} AU1_{3\text{ to }2} \xrightarrow{4\%;30\%} AU1_{2\text{ to }1} \xrightarrow{0\%;29\%} AU1_{1\text{ to }2} \xrightarrow{1\%;26\%} AU1_{1\text{ to }2} \xrightarrow{0\%;26\%} AU1_{2\text{ to }1} \xrightarrow{29\%;49\%} \text{fin parole}$$

Cette version de SMART permet ainsi de fournir une information multimodale indiquant comment utiliser tout un ensemble de signaux sociaux sur des tours de parole avec une information temporelle précise. Il est ainsi possible de choisir le nombre de co-occurences à sélectionner et les signaux qui doivent apparaître. Des règles communes à plusieurs états affectifs peuvent aussi être utilisées pour enrichir le rendu.

Une dernière limite est la mise en place effective de cette information pour la modification du comportement d'un agent. En effet, par rapport aux contraintes des normes BML et SSML, il s'est avéré assez compliqué de coordonner efficacement les différents signaux. Les unités de temps peuvent être différentes : par exemple, un contour prosodique est mesuré en pourcentage là où une expression faciale est mesurée en seconde. De plus, un signal ne va pas forcément se désactiver lors d'une interaction mais osciller entre différentes valeurs ce qui est difficile à transcrire en *BML*.

Une solution, mise en place mais pas encore évaluée, considère les valeurs trouvées dans les règles d'associations temporelles d'un signal comme des points de passage à atteindre. Elles servent alors à interpoler une trajectoire pour les valeurs du signal lors du tour de parole. Cela présente l'avantage de pouvoir combiner facilement plusieurs règles co-occurrentes d'un même signal. En considérant les informations obtenues dans le calcul d'une règle d'associations comme un point de passage et en utilisant les informations de

co-occurences de règles, il est ainsi possible d'obtenir des variations complexes pour un signal comme cela est illustré dans la figure 4.12.

En utilisant ensuite ces trajectoires avec le même principe qui a été utilisé pour créer l'agent-miroir lors de l'élaboration du corpus *POTUS*, un *morphing* du signal étudié va pouvoir modifier le comportement de l'agent. Il est ainsi possible de faire varier ses expressions faciales, mouvements de tête ou sa prosodie tout en conservant les mouvements de sa bouche pour garder la synchronisation avec sa voix, comme cela est illustré dans la figure 4.13. Cela permet d'avoir une vidéo de l'agent, avec le contenu verbal et la production de la parole originale, mais également avec des expressions faciales, mouvements de tête et f_0 correspondant à un état affectif voulu. Cette méthode est en fin de développement pour une évaluation future.

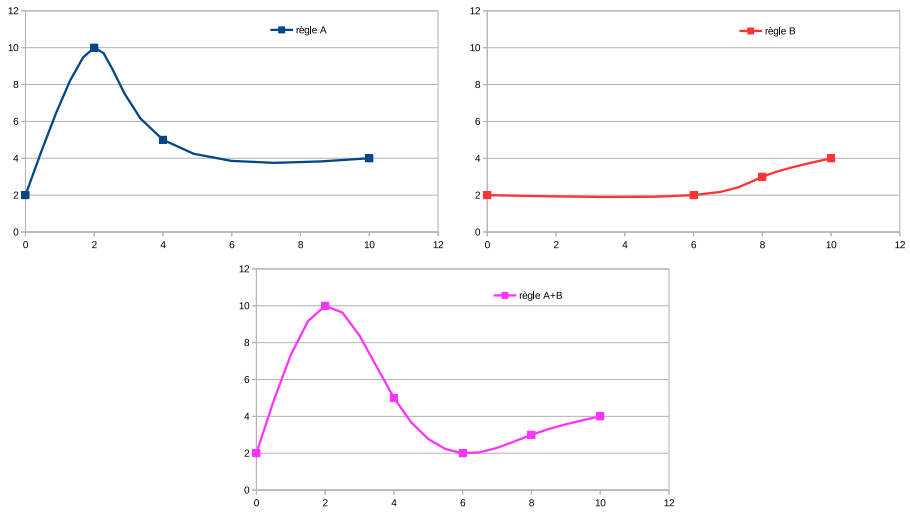

FIGURE 4.12 – Utilisation des règles pour l'extraction de points de passage pour un signal social. La combinaison de règles différentes en fonction de leurs co-occurences permet d'obtenir des variations de signaux plus complexes pour la synthèse.

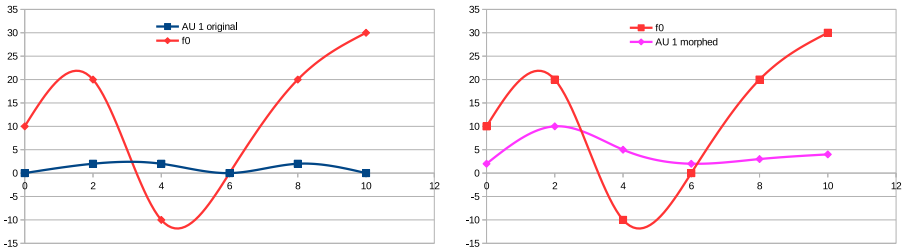

FIGURE 4.13 – Illustration du *morphing* possible pour transformer des données originales afin d'exprimer une attitude. Action Unit 1 est ici modifiée pour suivre les informations des règles exprimant une attitude voulue et correspondant à la forme de la f_0 détectée.

4.5 Conclusion

Ce chapitre a permis d'explorer une solution utilisant la fouille de données pour effectuer de la synthèse de comportements chez un agent conversationnel. Les résultats des premières études soulignent l'intérêt de cette approche et ont été validés lors d'évaluations perceptives. *SMART* a ainsi montré son utilité pour trouver de l'information temporelle précise en liant des signaux sociaux à l'expression d'une attitude sociale pour pouvoir ensuite être synthétisée sur un agent virtuel.

La multi-modalité s'est révélée plus compliquée à intégrer mais l'étude a souligné l'intérêt d'une approche *fusion tardive*. Cette dernière a été mise en place et est en cours de finalisation. Elle considère l'information trouvée comme une trajectoire, des points de contrôle à atteindre pour l'expression de l'état affectif désiré. Elle devrait permettre d'évaluer l'influence des différents signaux dans l'expression d'attitudes dans le temps. Cela montre aussi une limite de la norme *SAIBA* qui se révèle trop rigide pour ces transformations.

Ce qu'il faut retenir :

Contributions :

- une méthode de fouille de données qui permet de trouver des associations de signaux sociaux dans le temps sous forme de règles
- des scores qui permettent de lier ces règles à un phénomène affectif, ici les attitudes sociales, mais également l'utilisation d'un coefficient qui garantit leur généralisation entre les individus
- une adaptation de ces règles pour la synthèse d'attitudes sociales

Chapitre 5

Un modèle d'apprentissage profond : le système SSN

Sommaire

Ce chapitre présente les travaux effectués lors d'une collaboration au sein de l'*Institute of Creative Technologies* (*University of South California*, Los Angeles) durant l'été 2017. Il s'agit d'explorer les dernières possibilités offertes par l'évolution de l'apprentissage profond ces dernières années. La combinaison des techniques en recherche de séquences avec les dernières avancées en séparation de domaines et apprentissage multi-tâches ont permis l'élaboration dans cette thèse d'un modèle d'analyse simultanée de plusieurs états affectifs. Les premiers résultats et les développements futurs sont détaillés ici.

5.1 L'utilisation de réseau de neurones artificiels

5.1.1 L'apprentissage profond

L'apprentissage profond est une classe de méthodes d'apprentissage automatique, initiée dès les années 80 (Rumelhart et al., 1985; LeCun et al., 1989), mais dont l'utilisation s'est réellement développée à partir de 2012. Le principe est d'entraîner des modèles computationnels en utilisant différentes couches de traitements afin d'apprendre des représentations des données avec plusieurs niveaux d'abstractions. Chaque couche prend en entrée la sortie de la précédente. Une couche va extraire et transformer des caractéristiques dans les données qui lui sont fournies. La propagation entre les différentes couches va ensuite permettre d'établir une hiérarchie entre ces caractéristiques et ainsi en proposer différents niveaux d'abstraction.

Cette capacité à extraire et hiérarchiser des caractéristiques automatiquement a permis de nombreux progrès dans le domaine de la vision par ordinateur ou du traitement de la parole. En effet, avec un nombre conséquent de données et des moyens de calcul appropriés, la représentation ainsi trouvée et une bonne capacité d'analyse vont permettre d'améliorer les tâches d'apprentissage (Baydin et al., 2015). La reconnaissance de formes en est un bon champ d'application. Dans un système classique, deux éléments étaient nécessaires : un extracteur de caractéristiques et un classificateur entraînable. L'extracteur est défini à la main, souvent empiriquement, afin de transformer les données en entrées (un tableau de pixel par exemple) en un vecteur de caractéristiques. Ce vecteur va ainsi décrire la présence ou l'absence d'un certain nombre de motifs dans les données observées. Le classificateur va assigner un poids à chacune des caractéristiques calculées et ainsi, selon la comparaison à un seuil, déterminer la classe la plus probable pour les données. Ce sont ces poids qui sont appris lors de l'apprentissage et ils forment une représentation prototypique de la classe à laquelle le vecteur de caractéristiques est ensuite comparé. L'inconvénient de cette méthode est que la définition des caractéristiques pertinentes n'est pas toujours facile à réaliser. Cela peut demander beaucoup de temps et d'expertise humaine pour un résultat qui ne sera pas facilement transposable à une problématique légèrement différente.

L'apprentissage profond va proposer une alternative en entraînant globalement le système qui est composé d'une série de modules, de couches successives. En fait, chaque module est entraînable avec un jeu de paramètres à ajuster. A chaque étape de l'apprentissage, le système va ajuster ces paramètres afin de rapprocher la sortie globale du système de la cible désirée. Un gradient permet d'estimer l'erreur et sa variation pour chaque paramètre modifié. En le rétro-propageant entre toutes les couches, il indique ainsi si l'erreur augmente ou diminue. Cela permet au système de s'ajuster pour approcher une meilleure

solution à l'étape d'entraînement suivante. Un exemple de cette architecture en couches est visible dans l'image 5.1 qui montre l'intérêt de la représentation interne pour trouver la sortie pertinente pour toute entrée.

5.1.2 Les réseaux de neurones récurrents

Un réseau de neurones récurrents est un réseau de neurones artificiels présentant au moins un cycle. Cela signifie qu'il existe au moins un neurone dont la sortie sera une de ses propres entrées, après un éventuel traitement par un ou plusieurs autres neurones. Ils sont ainsi très répandus dans des applications de reconnaissance automatique de la parole ou de l'écriture manuscrite : la mémorisation des informations précédentes va leur

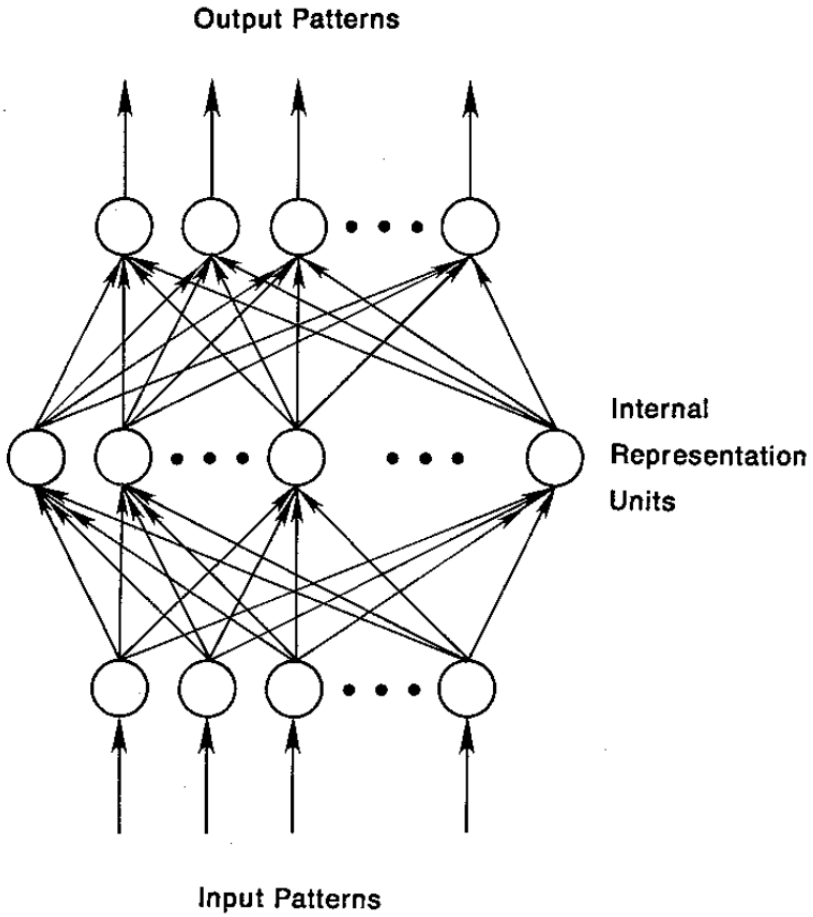

FIGURE 5.1 – Image illustrant le fonctionnement d'un réseau multicouches, issue de Rumelhart et al. (1985). L'information d'entrée est ré-encodée dans une représentation interne, cachée, à partir de laquelle la sortie va être générée à son tour. Si le nombre d'unité dans la représentation cachée est suffisant, toute entrée pourra être encodée afin de trouver la sortie appropriée.

permettre de différencier un phonème d'un autre par exemple. L'idée sous-jacente est que cette boucle va leur permettre de conserver de l'information en mémoire et la prendre en compte dans l'apprentissage. Ainsi, on peut visualiser un réseau de neurones récurrents comme un réseau classique mais constitué d'une multitude de copies du même réseau (voir fig.5.2). Cette structure en chaîne indique leur appétence à traiter des problématiques sur des séquences ou des listes d'événements.

Ainsi, les réseaux de neurones récurrents sont censés pouvoir conserver l'information passée afin de prendre leur décision. S'il s'agit de prendre l'information récente en compte, ces modèles pourront être efficaces. Ainsi, un modèle de langage bien entraîné, sur cette thèse par exemple, qui prend en entrée un morceau de phrase et la complète, réussira à prédire pour l'entrée "Agent Conversationnel" le mot *Animé*. Cependant, il peut y avoir des situations où plus de contexte est nécessaire pour obtenir un résultat satisfaisant. Si l'on considère l'entrée "J'ai grandi en France [...]. Je parle couramment", la réponse attendue est *français*. Un réseau de neurones récurrents aura la capacité, grâce à l'information récente, de comprendre qu'un nom de langage est attendu. Cependant, le contexte du pays peut être trop éloigné dans le texte pour que le réseau puisse y accéder. Cela est lié au problème de disparition du gradient qui fait que les poids des événements passés diminuent de manière exponentielle.

Une solution couramment proposée est l'utilisation de réseau *Long short-term memory* (LSTM) introduit par Hochreiter and Schmidhuber (1997). Tout comme les réseaux de neurones récurrents classiques, les LSTM ont également une structure en chaîne. Cependant, comme cela est visible dans la figure 5.3, les couches internes qui les composent diffèrent : les 4 éléments vont interagir afin de pouvoir conserver simultanément de l'information récente et distante tout en les mettant à jour lorsque cela est nécessaire. L'élément principal concerne l'état de la cellule qui va être propagé tout au long de la chaîne.

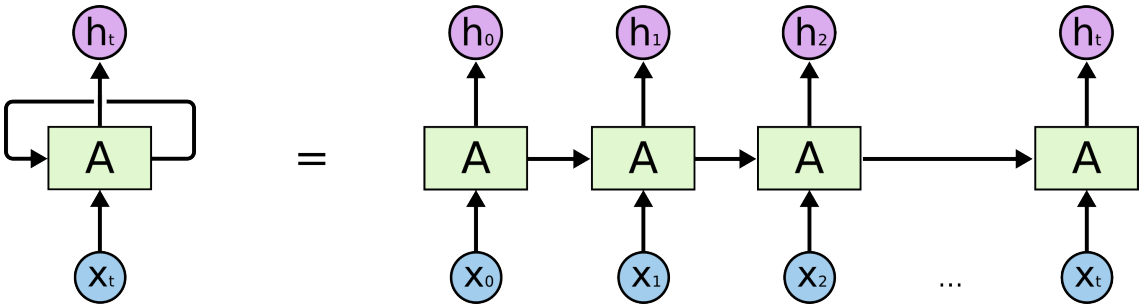

FIGURE 5.2 – Un réseau de neurones récurrents avec sa boucle caractéristique et sa visualisation "déroulé"

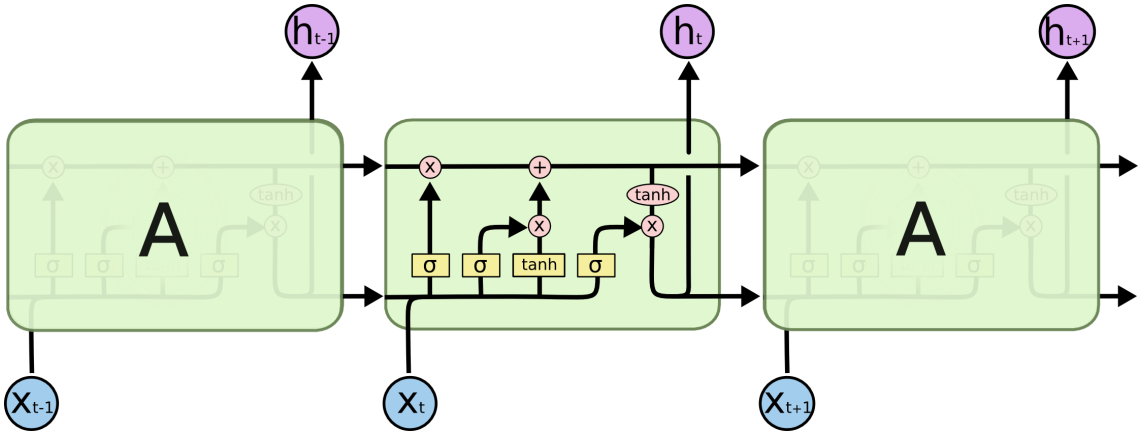

FIGURE 5.3 – Schéma de fonctionnement d'un LSTM

Les LSTM ont la possibilité de modifier cet état grâce à trois portes distinctes. La première porte va décider de ce qui est à oublier ou à conserver des états précédents. Dans l'exemple de complétion de phrases, l'information peut concerner le sujet courant comme son genre afin de choisir correctement le pronom correspondant. Lors de la rencontre d'un nouveau sujet, cette porte permet d'oublier l'ancien sujet. La deuxième porte va ensuite permettre d'effectuer la mise à jour des informations nécessaires. C'est là, dans notre exemple, que le genre du nouveau sujet sera mis à jour. La troisième porte va s'occuper de la sortie à fournir au réseau en utilisant l'état de la cellule et les informations en entrée. Dans notre exemple, il peut indiquer si le sujet est singulier ou pluriel, masculin ou féminin, ...

Des variantes du modèle LSTM existent comme les *Gated Reccurent Unit* (Cho et al., 2014) par exemple, mais, dans le cadre de cette thèse, c'est ce modèle classique qui a été utilisé pour servir de base aux modèles proposés. Il permet ainsi de prédire à partir d'une séquence en entrée la configuration au pas de temps suivant. Un modèle composé d'un simple LSTM servira comme un système de référence (*baseline*) et les améliorations apportées par la contribution de cette thèse, le *Social Separation Network*, y seront comparées.

5.1.3 Adaptation de domaines et séparation de domaines

L'*Adaptation de domaines* (*Domain Adaptation*) est un cas particulier de l'apprentissage par transfert. L'idée est d'effectuer une tâche d'adaptation d'un domaine source vers un domaine cible. Un exemple d'application, pour des images, serait d'avoir un corpus d'images de sacs à dos dans lequel les photos ont été faites en studio et un deuxième contenant également des sacs à dos, mais en extérieur, dans un contexte réel. Le but de l'adaptation de domaines sera de trouver ce qui est commun aux deux corpus pour par exemple recon-

naître les sacs dans le corpus "*in the wild*". Le but est donc de construire une représentation commune, partagée, des deux domaines de manière plus ou moins supervisée.

Csurka (2017) propose une revue très complète des applications de l'adaptation de domaines en vision. Elle y explique en particulier l'intérêt de son utilisation pour l'annotation automatique de nouvelles données à partir d'un corpus source connu. Elle souligne en particulier l'intérêt de l'apprentissage multi-tâches qui consiste à apprendre simultanément différentes tâches en représentant l'information commune à celle-ci, améliorant au passage les performances. Caruana (1998) indique ainsi que ces approches augmentent la généralisation des résultats en discernant ce qui est propre à une tâche par rapport aux autres.

Parmi toutes ces approches d'adaptation de domaines, le DANN (*Domain-Adversarial Neural Network*, Ganin et al. (2016), introduit l'idée d'utiliser une couche d'inversion du gradient (*GRL*) dans son architecture. Le *GRL* va modifier le gradient lors de l'étape de rétro-propagation en le multipliant par -1.
Un exemple d'utilisation est visible dans la figure 5.4 qui présente un réseau d'adaptation de domaines lors d'une tâche de classification. A partir d'une entrée x, un extracteur de caractéristiques et un classifieur de labels forment une architecture "classique" d'un système de reconnaissance. Cependant, une couche *GRL* est également placée à la sortie de l'extracteur de caractéristiques avant de passer dans un classifieur de domaines. Le système minimise simultanément les fonctions objectif de prédiction des labels et des domaines.

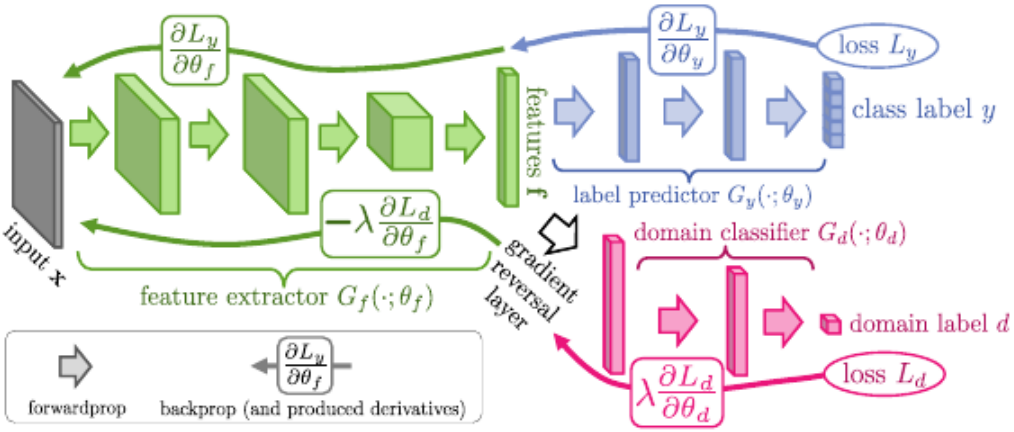

FIGURE 5.4 – L'architecture du DANN telle que présentée par Ganin et al. (2016). On y voit l'extracteur de caractéristiques (vert) et un classifieur de classes (bleu), formant une architecture "classique". L'utilisation d'un *GRL* et d'un classifieur de domaines (rose) assure que les caractéristiques utilisées par le classifieur de classes sont bien indépendantes du domaine.

La couche *GRL* va alors assurer que les caractéristiques extraites automatiquement pour déterminer les classes ne sont plus discriminantes entre les domaines. Les caractéristiques utilisées par le classifieur de classes (en bleu dans la figure) sont alors indépendantes du domaine initial et peuvent donc être généralisées.

Bousmalis et al. (2016) va améliorer cette idée en développant l'architecture du *Domain Separation Network*, illustrée dans la figure 5.5. Ils proposent la notion d'un sous-espace privé à chaque domaine ainsi qu'un sous-espace partagé. Le sous-espace privé va capturer des propriétés propres à chaque domaine, comme l'arrière-plan ou des caractéristiques bas niveau dans l'image. L'espace partagé, lui, va capturer les caractéristiques communes aux deux domaines, par exemple la forme ronde d'un sac à dos et la notion de bretelles. Pour cela, des fonctions objectif vont assurer une bonne séparation entre les représentations privées et partagées, mais aussi garantir que la somme de ces représentations garde une bonne reconstruction par rapport à la donnée d'origine.
Visibles dans la fig.5.5, les fonctions objectif $\mathcal{L}_{\text{différence}}$ assurent l'orthogonalité des représentations trouvées entre l'espace privé et l'espace partagé. La fonction $\mathcal{L}_{\text{similarité}}$ utilise un GRL pour assurer que la représentation trouvée par l'encodeur partagé ne permettent plus de retrouver le domaine d'origine. Un classifieur avec la fonction objectif $\mathcal{L}_{\text{classe}}$ assure néanmoins que les caractéristiques conservées permettent toujours de reconnaître la classe de l'entrée. C'est une utilisation similaire au *DANN* de Ganin et al. (2016). Les fonctions objectif $\mathcal{L}_{\text{recon}}$ assure qu'il n'y a pas eu de pertes d'informations en reconstruisant les entrées à partir des représentations communes et privées de chaque domaine. Ainsi, chaque "*encodeur privé*" contient une représentation propre à chaque domaine là où l'"*encodeur*

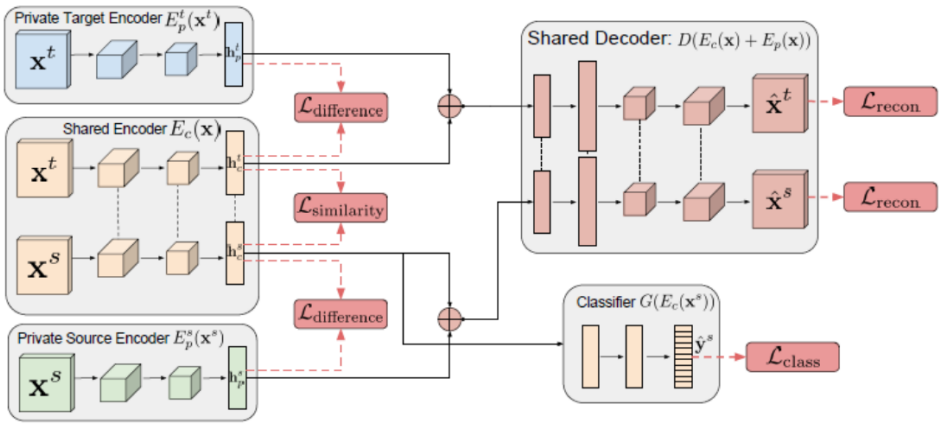

FIGURE 5.5 – L'architecture du *Domain Separation Network* telle que présentée par Ganin et al. (2016). Des encodeurs privés et partagés vont permettre de construire des représentations communes et spécifiques pour chaque domaine. Un décodeur commun va permettre l'apprentissage en reconstruisant les entrées proposées.

partagé" possède une représentation commune aux deux et la somme de chaque représentation permet une reconstruction fidèle des données. Ils démontrent l'efficacité de cette technique sur plusieurs jeux de données en reconnaissance d'images et donc assigner un label entre différents domaines.

5.2 Le modèle du *Social Separation Network*

5.2.1 Introduction

Le modèle développé dans le cadre de cette thèse s'inspire des différentes architectures qui ont été présentées dans la partie précédente. Avec le *Domain Separation Network*, Ganin et al. (2016) introduisent les idées d'espaces privés et d'espace partagé entre deux domaines. Ce concept est repris dans le développement du *Social Separation Network* avec un focus sur la dynamique des signaux. En effet, les théories en psychologie et sociologie(cf tableau 1.1) montrent que les modulations dans le temps de différents signaux sociaux permettent l'expression de différents phénomènes affectifs.

Le *Social Separation Network* modélise cet aspect dynamique grâce à des couches *LSTM* afin de trouver des représentations temporelles dans des signaux sociaux extraits automatiquement. Les idées de domaines privés et partagé sont appliquées aux différents phénomènes affectifs. Les représentations propres à chacun ainsi que la représentation commune peuvent ensuite être synthétisées sur un agent virtuel. Cela permet de visualiser la façon dont les signaux sociaux sont utilisés dynamiquement pour chaque cas. Ce travail pourra être adapté pour synthétiser directement les phénomènes affectifs ensuite.

5.2.2 Présentation du modèle

Cette section détaille le modèle du *Social Separation Network* qui a été développé dans cette thèse. Il est conçu pour être appliqué à l'étude de signaux sociaux lors de l'expression de deux tâches notées tâche_1 et tâche_2. Ces tâches peuvent correspondre à des phénomènes affectifs (attitude, personnalité, ...) ou être plus générales comme le jeu d'un acteur ou le genre d'une personne. Ces différentes possibilités sont illustrées dans les études, présentées dans la section5.3.
Le *Social Separation Network* cherche à trouver dans la dynamique des signaux sociaux des représentations propres pour chaque tâche et une commune aux deux. Son architecture, illustrée dans la figure 5.6, s'inspire du *Domain Separation Network* afin de trouver ces représentations avec quatre grandes parties dans son architecture : un encodeur/classifieur pour chaque état affectif, un encodeur partagé et un décodeur pour la reconstruction, mais

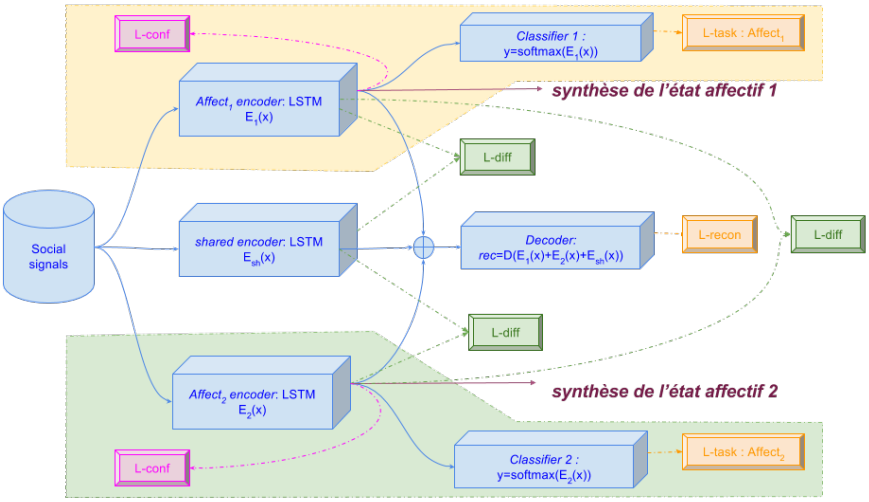

FIGURE 5.6 – Le modèle du Social Separation Network

utilise des *LSTM* pour prendre en compte la temporalité.

Le *Social Separation Network* prend en entrée, notée x, un ensemble de signaux sociaux extraits automatiquement sur une fenêtre composée de w images contiguës issues du corpus vidéo étudié. Cette entrée est accompagnée de deux vecteurs indiquant les valeurs de chaque tâche. Pour cela, l'encodage *one-hot* a été utilisé : il consiste à représenter les différents états possibles de chaque tâche avec un vecteur où un seul bit est à 1 et tous les autres à 0.
Par exemple, pour une tâche avec m valeurs possibles, un vecteur de taille m est utilisé. Une première valeur sera encodée en mettant la 1^{re} case à 1 et toutes les autres à 0, pour la seconde valeur, la 2^{me} case sera à 1 et toutes les autres à 0, ... Bien que gourmande en mémoire, cette représentation présente l'avantage d'avoir un coût constant pour accéder et modifier un état.

Pour chaque tâche, une partie est dédiée à trouver une représentation privée qui permettra de le reconnaître grâce à un pipeline composé d'un encodeur *LSTM* et d'un classifieur *Softmax*. A partir de l'entrée x, l'encodeur va trouver une représentation pour la tâche$_1$ (resp. tâche$_2$), notée $E_1(x)$ (resp. $E_2(x)$). L'utilisation de la fonction *Softmax* va ensuite déterminer la probabilité de chaque classe de l'affect étudié, notée $\hat{y}_1 = G(E_1(x))$ (resp. $\hat{y}_2 = G(E_2(x))$). La fonction objectif associée cherchera à diminuer la fonction de vraisemblance logarithmique inverse (négative *log-likelihood*) pour chaque cas avec la formule 5.1.

$$\mathcal{L}_{\text{task}} = -\sum_{i=0}^{m} y_i.log(\hat{y}_i) \tag{5.1}$$

Un troisième encodeur *LSTM* va trouver une représentation commune aux deux tâches, notée $E_{sh}(x)$. Un décodeur va ensuite utiliser ces trois représentations pour reconstruire le signal en les ajoutant : $\hat{x} = E_{sh}(x) + E_1(x) + E_2(x)$. Pour cela, la fonction objectif associée va chercher à réduire l'erreur quadratique moyenne entre l'entrée et la reconstruction, présentée dans la formule 5.2.

$$\mathcal{L}_{\text{rec}} = \sqrt{\sum_{i=0}^{m} \frac{(\hat{x}_i - x_i)^2}{m}} \tag{5.2}$$

La séparation des domaines est assurée grâce à deux types de fonctions objectif, notée $\mathcal{L}_{\text{dif}}$ et $\mathcal{L}_{\text{conf}}$.
La première, $\mathcal{L}_{\text{dif}}$, va encourager les sorties de deux encodeurs à être bien orthogonales, encourageant ainsi les représentations trouvées à indiquer différents aspects de l'entrée. Pour cela, le produit scalaire entre ces encodages est calculé et la fonction cherche à minimiser ce résultat.
La seconde fonction, $\mathcal{L}_{\text{conf}}$, est une adaptation de la *similarity loss* du *Domain Separation Network*. Le but est d'assurer que les représentations des encodeurs privée d'une tâche ne puisse plus reconnaître l'autre tâche. Comme dans Ganin et al. (2016) (figure 5.4), une couche d'inversion du gradient et un classifieur sont utilisés pour assurer que les caractéristiques trouvées par encodeur de la tâche_1 (resp. tâche_2) ne permettent plus de reconnaître la tâche_2 (resp. tâche_1).

$$\mathcal{L}_{\text{conf}} = -\sum_{i=0}^{m} y_i.log(\hat{y}_i) + (1 - y_i).log(1 - \hat{y}_i) \tag{5.3}$$

La combinaison de $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{dif}}$ et $\mathcal{L}_{\text{conf}}$ assure alors que la représentation à la sortie de chaque encodeur privé est bien spécifique à une tâche et indépendante de l'autre tâche. Il va ainsi être possible de trouver ce qui est commun à l'expression d'attitude (tâche_1) entre différents acteurs (jeu d'acteur : tâche_2). Le *Social Separation Network* peut également se confronter à la théorie, par exemple en prenant pour chaque variable une des dimensions du circomplexe d'Argyle (cf figure 1.4).

5.3 Évaluations

Afin d'illustrer l'intérêt de cette méthode, celle-ci a été appliquée sur les deux corpus retenus dans cette thèse : *Semaine-db* et le corpus *POTUS*. Les premiers résultats sont encourageants pour sa validation et montre son intérêt pour de la classification et l'utili-

sation d'agents virtuels permet de visualiser les représentations trouvées. Son adaptation à la synthèse est la prochaine étape dans sa réalisation.

Le mode opératoire des deux études est similaire mais elles étudient chacune un corpus et deux tâches différentes. Dans les deux cas, les signaux sociaux étudiés sont des *action units* extraites automatiquement comme cela a été détaillé dans la partie 3.2. Les données sont ensuite séparées en trois jeux : un jeu d'entraînement, un jeu de validation et un jeu de test, ce dernier étant complètement indépendant des deux autres i.e. ne provenant des mêmes fichiers vidéos. Le modèle est donc entraîné et validé sur les deux premiers jeux avant d'être testé sur le dernier.
A chaque fois, un modèle de référence, intitulé *baseline* dans la suite de ce manuscrit, servira de comparaison. Il s'agit d'un classifieur de séquences classique composée d'un *LSTM* pour prendre en compte la dynamique et d'une couche *dense* utilisant la fonction *softmax* pour indiquer la probabilité d'appartenir à une classe. Dans la figure 5.6, la *baseline* se retrouve dans la partie privée de chaque tâche avec une seule fonction objectif : $\mathcal{L}_{\text{rec}}$. En effet, le but ici n'est pas de trouver les paramètres optimaux pour les couches *LSTM* et de classification mais d'évaluer l'apport de l'architecture *SSN* par rapport à la *baseline*.

Lors de test, différentes configurations d'entraînement ont ensuite été testées avec des variations de :

- la taille des séries temporelles en entrée
- le nombre d'*epochs* correspond au nombre de passages sur l'ensemble des données d'entraînement pour apprendre le modèle
- le *batch size* correspond au nombre d'exemples d'entraînement utilisés à chaque propagation pour estimer les paramètres du modèle

Quatre mesures ont été faites pour observer les différences des deux approches : le score, la précision, le rappel et le F-score. Le score est le taux de bonne classification. La précision indique la capacité du classifieur à ne pas classer comme positif un échantillon qui est négatif. Le rappel indique la capacité du classifieur à trouver tous les échantillons positifs. Le F-score est la moyenne harmonique de la précision et du rappel. Chaque étude présentera ses résultats sur différentes fenêtres de temps, mesurées en nombre d'images concomitantes.

Il est ensuite possible de visualiser les représentations privées pour chaque tâche ainsi que la représentation partagée. En effet, E_1, E_2 et E_{sh} contiennent les valeurs des *action units* que chaque encodeur considère comme pertinentes. Pour chaque étude, des vidéos sont disponibles où sont synthétisées sur l'agent, de haut en bas, de gauche à droite, les valeurs originales, la représentation partagée (E_{sh}), la représentation privée de la tâche$_1$ (E_1) et celle de la tâche$_2$ (E_2).

5.3.1 Études 1 : attitude et jeu d'acteur

Le modèle a tout d'abord été utilisé avec *Semaine-db*. Les signaux sociaux en entrées étaient les action units et les mouvements de têtes extraits automatiquement. Les deux domaines séparés par le modèle ont été définis comme suit. Le premier est une approximation de l'attitude exprimée, comme cela a été fait avec SMART, avec le personnage joué : Poppy, l'amical, et Spike, l'hostile. Le second cherche à modéliser tout ce qui correspond au jeu d'acteur et cherche donc à différencier les opérateurs humains qui ont été identifiés : Cowie, McKeown, ... Il s'agit donc d'utiliser le *Social Separation Network* pour établir une représentation privée de l'expression de l'amicalité et de l'hostilité (partie Poppy/Spike) en la différenciant de ce qui est propre aux acteurs.

Une première évaluation a consisté à observer l'influence de cette approche sur la reconnaissance des personnages joués. Elle a suivi le protocole présenté en introduction, et obtenu les résultats, présentés dans le tableau 5.1, sur différentes fenêtres de temps. Les différents résultats étant similaires, seul le F-score est présenté. *SSN* obtient sur tous les points de meilleurs scores que la *baseline*. De plus, il apprend également à reconnaître un autre phénomène affectif sur lequel il obtient des scores similaires. Le gain n'est pas forcément très important mais il souligne l'intérêt de l'architecture du *Social Separation Network* pour améliorer les résultats par rapport à un modèle basique.

Le modèle du *Social Separation Network* permet également de visualiser les représentations qui appartiennent à chaque domaine. Il est ainsi possible, à partir d'une entrée, d'observer ce qui est spécifique au personnage joué, à l'acteur et commun aux deux comme cela est visible dans ces vidéos [1] . Cette représentation accompagnée du score de classification, le *softmax* donnant la probabilité d'appartenir à une classe, devrait permettre le développement futur d'un système de synthèse. Il consisterait à utiliser les représentations ayant donné un score de reconnaissance satisfaisant (i.e. supérieur à un score voulu) pour entraîner un nouvel LSTM en mode "sequence to sequence". L'extension du système

taille en nombre	F-score modèle	
d'images considérées	baseline	SSN
3	0,66	0,68
5	0,66	0,67
10	0,67	0,74
30	0,67	0,75

TABLEAU 5.1 – Résultats des scores de reconnaissance sur Semaine-dB du *SSN* avec différentes fenêtres de temps

1. Poppy : https://youtu.be/cIU1QlMjOKA
Spike : https://youtu.be/_vibZ8HdKVo

pourra ainsi, dans un premier temps, apprendre les représentations privées et communes à deux états affectifs et, ensuite, les utiliser pour s'entraîner et ainsi générer de nouveaux comportements avec les affects voulus.

5.3.2 Étude 2 : les deux axes d'Argyle

En suivant exactement le même design d'expérimentations que pour l'étude 1, le corpus *Potus* a été également étudié avec le *Social Separation Network*. Les deux états affectifs servant de domaine étaient l'annotation en amicalité et celle en dominance.

Cette étude exploratoire visait surtout à observer ce que la représentation partagée de ces deux dimensions contient. Pour cela, les annotations selon chaque axe ont été divisées en deux classes pour l'apprentissage : forte amicalité (resp. dominance) et faible amicalité (resp. dominance). La théorie, bien que contestée, s'attendrait à un vide lié à leur orthogonalité supposée.

La figure 5.7 montre des scores corrects de reconnaissance pour chaque dimension. Il est intéressant d'observer que les scores du *Social Separation Network* sont généralement supérieurs à la *baseline* et surtout plus stables dans le temps. Une explication possible est que l'ajout d'information le rend plus résistant au sur-apprentissage et assure donc plus rapidement une meilleure efficacité.

Des vidéos présentant la synthèse de chaque représentation sont consultables ici [2]. La vidéo originale y est visible en même temps que les représentations privées et partagées

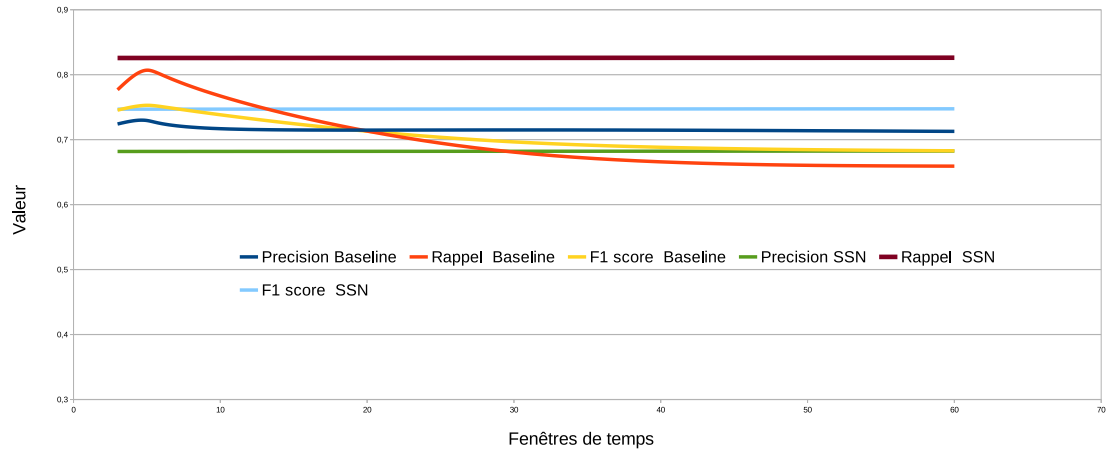

FIGURE 5.7 – Résultats des scores de reconnaissance sur *POTUS Corpus* du *Social Separation Network* sur différentes fenêtres de temps (en nombre d'images considérées)

2. https://youtu.be/uHUwCWVr9p0

de l'expression de l'amicalité et de la dominance, qui sont elles synthétisées sur l'agent Rodrigue. Il est intéressant de remarquer que la dominance se retrouve surtout dans les mouvements de tête et l'amicalité dans les expressions faciales. Enfin, la représentation privée est loin d'être vide, principalement composée d'expressions faciales. Cela peut s'expliquer par l'absence d'exemples négatifs dans l'expression de la dominance et de l'amicalité. Il serait néanmoins assez intéressant d'ajouter plus d'exemples pour confronter cette représentation à l'orthogonalité théorique de la représentation d'Argyle.

5.4 Conclusion

Ce chapitre a présenté les recherches effectuées lors d'une collaboration internationale de quelques mois. Elle a mené à l'élaboration d'un modèle d'analyse temporelle de signaux sociaux lors de l'expression de phénomènes affectifs.

Les deux évaluations ont montré l'intérêt d'utiliser les représentations privées et partagées lors de l'étude de la dynamique des signaux sociaux. Cette approche permet d'améliorer l'analyse et la reconnaissance tout en apprenant plusieurs tâches simultanément.

Cette information a pour vocation d'être utilisée lors d'une tâche de synthèse. Une solution simple consisterait à exploiter encore plus les informations du modèle. En effet, le classifieur *SOFTMAX* donne une probabilité de reconnaissance des différentes classes du phénomène ce qui peut servir de score sur la pertinence de la représentation trouvée par l'encodeur. Il devient alors possible de conserver les représentations satisfaisantes et d'entraîner un nouveau modèle constitué par exemple d'un simple *LSTM* pour apprendre à les utiliser pour exprimer une tâche voulue. En effectuant le même traitement pour la représentation partagée, la synthèse de comportement d'un agent est possible et serait intéressante à évaluer.

Il est à noter néanmoins que par rapport à l'approche *SMART*, les résultats présentés ici sont plus difficiles à interpréter. Les sorties du système sont bien moins interprétables pour un humain mais la possibilité de visualiser rapidement les représentations de chaque encodage aide à leur compréhension.

Il serait intéressant d'appliquer ce modèle à d'autres problématiques. Un développement en cours consiste à intégrer l'information dyadique en l'appliquant à Semaine-db avec une tâche concernant l'opérateur et l'autre tâche pour l'utilisateur.

Ce qu'il faut retenir :

Contributions :

- une méthode d'apprentissage profond qui permet d'analyser la dynamique de signaux sociaux et d'en extraire des représentations propres à différentes tâches mais également une représentation commune aux deux. Ces tâches peuvent être des phénomènes sociaux (attitudes : dominance et amicalité, par exemple), des informations contextuelles (jeu d'acteur).
- la capacité de synthétiser et visualiser ces représentations
- une amélioration des résultats de reconnaissance par rapport à un système de base simple, mais qui prend quand même en compte la dynamique

Chapitre 6

Conclusion

Sommaire

CETTE thèse s'est intéressée à la recherche d'informations temporelles qui lient les signaux sociaux lors de l'expression de différents états affectifs. Le cas d'application de l'étude de l'expression d'attitudes sociales a été retenu car la temporalité est un facteur essentiel dans ce phénomène.

Ce chapitre propose de résumer les différentes contributions, de discuter leurs limitations avant de proposer des solutions et des développements futurs pour les dépasser.

6.1 Résumé de la thèse

Le fil rouge de ces travaux est de s'intéresser aux différentes étapes de la méthodologie "classique" de construction de modèles pour l'animation d'agents conversationnels animés. Des corpus aux modèles en passant par l'extraction et le traitement des signaux sociaux, cette thèse propose des solutions pour trouver automatiquement cette information sur la dynamique de ces signaux en vue d'une tâche de synthèse.

Deux grandes parties ressortent de ces études. La première concerne les caractéristiques utilisées : elles ont été extraites automatiquement ce qui a mené également à l'élaboration d'un corpus de bonne qualité, annotés en attitudes sociales. La seconde traite du développement de modèles prenant en compte l'information temporelle dans l'expression de phénomènes affectifs qui permettent ensuite de synthétiser ces phénomènes avec un agent virtuel. Les principales contributions vont maintenant être rappelées ici avant d'aborder les perspectives à moyen et long termes de ces travaux.

6.1.1 Travail sur les signaux et annotations :

Un travail important a été mené sur l'extraction automatique de caractéristiques représentant les différents signaux sociaux afin de pouvoir les utiliser dans la synthèse d'un agent conversationnel animé. Les deux contraintes principales liées à l'extraction automatique et à la génération vont maintenant être rappelées ainsi que les solutions proposées.

Le choix d'une extraction automatique des signaux a été largement motivé par le contexte de cette thèse. Elle s'est déroulée au sein du labex SMART et les expertises des différents partenaires permettaient d'avoir accès à des savoirs et des outils aussi bien en vision (extraction des *action units* ou des mouvements de têtes) qu'en traitement de la voix (tour de parole et prosodie) (cf section 3.2).
Malgré l'efficacité des outils utilisés, les données extraites étaient généralement bruitées et ne pouvait pas être utilisées directement dans les modèles. La contrainte de génération s'est avérée profitable pour résoudre ce problème. En analysant des vidéos où l'agent effectue différents usages de ses *action units*, l'estimation de paramètres correctifs à appliquer aux valeurs du détecteur a permis de compenser ces erreurs (cf section 3.4.3).

La construction du corpus *POTUS* (voir section 3.4) a permis de valider ce recalage. Ce corpus sera bientôt disponible pour la communauté. Il propose un ensemble de vidéos de bonne qualité, des annotations précises en attitudes sociales et des caractéristiques extraites automatiquement. La moitié de ce corpus est composée d'allocutions de Barack Obama et l'autre moitié de vidéos où un agent virtuel clone l'ancien président. Cela permet d'évaluer l'impact de l'utilisation d'un agent virtuel par rapport à un humain.

Contributions de cette partie :

- un corpus d'allocutions annotés en attitudes sociales
- une méthode de traitement des signaux sociaux extraits automatiquement pour les recaler grâce à un agent virtuel

6.1.2 Des modèles pour l'étude de la dynamique :

Deux approches différentes ont été utilisées pour modéliser la dynamique des signaux sociaux lors de l'expression de phénomènes affectifs dans le temps. Au delà de la modélisation de ces variations, le but était également d'être discriminant entre deux ou plusieurs classes et de pouvoir utiliser l'information extraite dans une tâche de synthèse.

Le premier modèle, *SMART* (voir section 4.4), est basé sur de la fouille de données. Il recherche des règles d'associations temporelles entre les signaux transformés en événements symboliques.
Un premier travail consistait à symboliser ces signaux tout en conservant l'information nécessaire pour la tâche de synthèse. Cela implique d'avoir des regroupements suffisamment fins pour pouvoir les générer ensuite de manière précise, ce qui diffère d'une tâche de reconnaissance où un regroupement plus large peut néanmoins permettre une reconnaissance efficace.
SMART est inspiré de *TITARL*, l'algorithme de fouille de données initial, et permet de discriminer les résultats propres à un phénomène affectif mais aussi repérer ceux qui sont communs à tous. *SMART* assure aussi une bonne généralisation des résultats en assurant qu'ils sont partagés par plusieurs personnes et non pas spécifique à un individu.
Deux études mono-modales ont obtenus de bons résultats lors de la synthèse de ces règles pour faire exprimer à un agent différentes attitudes. L'étude multi-modale a permis de définir une bonne stratégie pour trouver de l'information permettant de mixer différents signaux dans le temps.

Le second modèle, *SSN*, est basé sur de l'apprentissage profond. L'intégration de réseaux *LSTM* pour modéliser la dynamique dans une structure d'apprentissage multi-tâches permet l'étude de plusieurs états affectifs en simultané.
Son avantage est de nécessiter très peu de pré-traitement sur les données et de pouvoir caractériser plusieurs classes d'apprentissage en même temps. En s'inspirant de techniques issues de la séparation de domaines, *SSN* permet de trouver des représentations privées et partagées entre deux phénomènes affectifs.
Les premières études sur la reconnaissance d'attitudes montrent que ses performances sont supérieures à un modèle de référence alors qu'il effectue un double apprentissage (i.e. il reconnaît deux phénomènes conjointement). La partie reconstruction de ce modèle permet d'observer les représentations caractéristiques trouvées. Des pistes pour adapter ces informations afin d'obtenir un modèle génératif sont également proposées.

Contributions de cette partie :

- le modèle *SMART* basée sur due la fouille de données. Il permet de trouver des associations temporelles entre des signaux sous forme de règles et de les adapter pour de la synthèse d'agents virtuels.
- Le modèle *SSN* qui utilise l'apprentissage profond pour trouver des représentations intéressantes dans la dynamique des signaux sociaux. il utilise les notions d'espaces privés et partagés pour analyser différents tâches dont les attitudes et montre une amélioration pour de la reconnaissance.

6.2 Perspectives à moyen terme

Une des principales difficultés rencontrée dans cette thèse tient dans l'évaluation des résultats multi-modaux pour la génération de comportements d'agents. Les travaux présentés ici ont voulu respecter les normes *BML/SSML* qui définissent *SAIBA*. Ces langages sont pratiques pour générer rapidement des comportements d'agents. Cependant, la validation des résultats obtenus dans cette thèse nécessite un contrôle fin des signaux, aussi bien dans leurs valeurs et dans leurs dynamiques. Même si des résultats ont été obtenus en mono-modalité, les premiers essais en multimodalité ont indiqué que le respect de ces normes n'était pas adapté pour valider les modèles. Deux raisons expliquent cette difficulté.

La première tient dans le formalisme *BML* : la gestion des *action units* se fait à partir d'un état désactivé et le *BML* va la modifier pour atteindre une valeur voulue à un instant donné avant de revenir au repos au bout d'un temps déterminé. L'exemple présenté dans la figure 6.1 [1] décrit le formalisme d'une activation de l'*action unit 1* pendant 4 secondes avec une intensité de 0.8. Ainsi AU_1 aura une intensité nulle au début, va augmenter jusqu'à atteindre 0.8 avant de redevenir nulle 4 secondes plus tard. Il n'y a pas de contrôle

```
<bml xmlns="http://www.bml-initiative.org/bml/bml-1.0"
     xmlns:ext="http://www.bml-initiative.org/bml/coreextensions-1.0"
         character="Alice"
         id="bml1">
    <face id="behavior1" amount="0.8" start="0" end="4">
      <ext:facs au="1" side="BOTH"/>
      <lexeme lexeme="WIDEN_EYES"/>
    </face>
</bml>
```

FIGURE 6.1 – Exemple de contrôle de l'AU4 (haussement de sourcils pendant 4 secondes) avec la norme *BML*

1. issue de http://www.mindmakers.org/projects/bml-1-0/

possible sur la façon d'atteindre ce pic à 0.8 ni de modifier la valeur par défaut de l'*action unit* avant et après.
Les résultats obtenus, aussi bien avec *SMART* qu'avec *SSN* ne trouvait pas nécessairement ces retours à des positions de repos. Il serait possible de faire évoluer *SMART* pour prendre en compte cette contrainte supplémentaire en le forçant à trouver des règles allant d'une position de repos à un retour à cette position de repos. Cependant, les résultats trouvées dans les études, en particulier du corpus *POTUS*, indiquent que le retour à cette désactivation n'existe pas forcément. Au contraire, plusieurs signaux vont rester activés pendant l'interaction et osciller autour d'une valeur.
Une solution intéressante pour ce problème serait d'intégrer des valeurs par défaut dans la norme *BML* ou de pouvoir intégrer plusieurs points de passage lors de ce contrôle. Il serait ainsi possible d'interpoler les intensités des *action units* ou des mouvements de tête directement avec le *BML* ainsi qu'intégrer une valeur par défaut qui ne soit pas forcément nulle.

La deuxième difficulté provient de l'échelle de temps en soi. Avec *SMART*, des résultats intéressants ont été trouvés avec les expressions faciales en secondes, et avec les contours prosodiques en pourcentage de réalisation de la phrase. Les études suivantes ont cherchées des associations multi-modale entre les *action units* et la f_0 aussi bien en seconde qu'en pourcentage. Cette dernière unité a donné de meilleurs résultats comme cela a été présenté.
Comme pour l'étude de la prosodie, la solution trouvée est de passer par du *morphing* plutôt que de la génération. L'idée est d'utiliser une vidéo et son son et d'en modifier les caractéristiques (f_0, valeurs d'*action units*, ...). Cela permet d'avoir un contrôle fin du temps. Cette solution évite aussi des problèmes comme la synchronisation des lèvres avec la production de la voix. La solution de *morphing* a donc été implémentée et les vidéos sont en cours de réalisation.
Une perspective d'étude intéressante consisterait à évaluer alors l'influence des différentes règles lors de la synthèse. Lors de l'étape de *morphing*, il est possible d'utiliser n règles d'un signal social, m règles d'un autre signal, etc, et cela pour toute les modalités. Avec une étude perceptive, il serait possible d'évaluer le poids de chacun des signaux dans l'expression d'attitudes chez un agent.

L'un des enseignements latent à cette difficulté est que les solutions proposées sont plus appropriées pour modifier de l'existant et non faire une synthèse complète. *SMART* trouve des points de passage pour chaque signal pour exprimer une attitude. *SSN* indique quelle évolution suivre à partir des positions précédentes pour se conformer à la représentation d'une attitude. Ces deux modèles indiquent comment modifier le comportement pour le "colorer" d'une attitude, en écho à la définition suivie.

Une deuxième limitation importante dans les travaux présentés ici est de s'être limité à l'intra-synchronie là où une attitude se développe lors d'une interaction. Cela s'explique par la volonté de trouver de l'information temporelle avant tout et cette limite a été surmontée en étudiant des corpus particuliers : dans *SAL SEMAINE-Db*, les intervenants jouent des caractères caricaturaux, dans le corpus *POTUS*, il s'agit d'allocutions. Ainsi, dans les deux cas, le retour de l'inter-actant ne modifiait pas l'attitude exprimée. Cela a permis de valider l'approche proposée et l'information temporelle trouvée.
Néanmoins, la prise en compte de l'inter-synchronie est nécessaire et les adaptations à faire sont détaillées dans les perspectives (section 6.3).

6.3 Perspectives long terme

Une information complémentaire serait de prendre en compte le comportement des autres intervenants de l'interaction. Les travaux précédemment effectués, en particulier Pecune et al. (2016b), indiquent que cela devrait améliorer les modèles.
SMART est déjà en état de prendre en compte cette information : il suffit juste d'ajouter l'information des autres intervenants dans le calcul des règles. Cependant, il serait intéressant de comparer, comme pour la multi-modalité, si les règles sont plus pertinentes en contenant des signaux de plusieurs intervenants ou si l'information est plus consistante en utilisant des règles d'intra-synchronie et leurs co-occurences.
Un modèle dyadique de *SSN* est en réflexion : il s'agirait d'utiliser une représentation partagée entre deux intervenants et d'avoir une représentation privée pour chacun. Ce modèle permettrait d'évaluer l'évolution d'une interaction dans le temps.

La synthèse de ces résultats reste un point difficile. Il semble que l'information dynamique trouvée est plus appropriée pour modifier un comportement existant que pour le générer totalement. Les résultats trouvés lors des analyses étaient difficilement adaptable pour la norme *SAIBA* qui est utilisée dans la synthèse d'agents.
Une piste pour résoudre ce problème serait de trouver une sorte de dictionnaire entre les animations *bml* et les règles trouvées. En utilisant un système encodeur-décodeur sur des données synthétiques, il serait intéressant de voir si un système de transposition des résultats pourrait être trouvé.

Bibliographie

Aggarwal, C. C. and Han, J. (2014). *Frequent Pattern Mining*. Springer Publishing Company, Incorporated.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*.

Aigrain, J., Dapogny, A., Bailly, K., Dubuisson, S., Detyniecki, M., and Chetouani, M. (2016a). On leveraging crowdsourced data for automatic perceived stress detection. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*.

Aigrain, J., Spodenkiewicz, M., Dubuisson, S., Detyniecki, M., Cohen, D., and Chetouani, M. (2016b). Multimodal stress detection from multiple assessments. *IEEE Transactions on Affective Computing*.

Albrecht, K. (2006). Social intelligence. *Journal of Leadership Excellence*.

Ales, Z., Duplessis Dubuisson, G., Şerban, O., and Pauchet, A. (2012). A methodology to design human-like embodied conversational agents. In *Proceedings of the International Workshop on Human-Agent Interaction Design and Models*.

Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences : A meta-analysis. *Journal of Psychological Bulletin*.

Argyle, M. (1975). *Bodily communication*. Methuen Publishing Company.

Arroyo-Figueroa, G. and Sucar, L. E. (1999). A temporal bayesian network for diagnosis and prediction. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*.

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*.

Bailly, G., Mihoub, A., Wolf, C., and Elisei, F. (2015). Learning joint multimodal behaviors for face-to-face interaction : performance & properties of statistical models. In *Proceedings of the Human-Robot Interaction. Workshop on Behavior Coordination between Animals, Humans, and Robots*.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface : an open source facial behavior analysis toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

Bänziger, T. and Scherer, K. R. (2010). Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing : A sourcebook*.

Barbulescu, A., Ronfard, R., and Bailly, G. (2016). Characterization of audiovisual dramatic attitudes. In *Proceedings of Interspeech*.

Batal, I., V., H., Cooper, G. F., and Hauskrecht, M. (2013). A temporal pattern mining approach for classifying electronic health record data. *ACM Transactions on Intelligent Systems and Technology*.

Bawden, R., Clavel, C., and Landragin, F. (2015). Towards the generation of dialogue acts in socio-affective ecas : a corpus-based prosodic analysis. *Journal of Language Resources and Evaluation*.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2015). Automatic differentiation in machine learning : a survey. *arXiv preprint arXiv :1502.05767*.

Boersma, P. and Weenink, D. (2017). Praat : doing phonetics by computer [computer program]. version 6.0.27.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. (2016). Domain separation networks. In *Proceedings of the Advances in Neural Information Processing Systems*.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Univ of California Press.

Burgoon, J. K., Buller, D. B., Hale, J. L., and Turck, M. A. (1984). Relational messages associated with nonverbal behaviors. *Journal of Human Communication Research*.

Burgoon, J. K. and Le Poire, B. A. (1999). Nonverbal cues and interpersonal judgments : Participant and observer perceptions of intimacy, dominance, composure, and formality. *Journal of Communications Monographs*.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., C., J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap : Interactive emotional dyadic motion capture database. *Journal of Language resources and evaluation*.

Cafaro, A., Vilhjálmsson, H. H., and Bickmore, T. (2016). First impressions in human–agent virtual encounters. *ACM Transactions on Computer-Human Interaction*.

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., and Valgardsson, G. S. (2012). First impressions : Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Proceedings of the International Conference on Intelligent Virtual Agents*.

Carney, D. R., Hall, J. A., and LeBeau, L. S. (2005). Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*.

Caruana, R. (1998). Multitask learning. In *Learning to learn*. Springer.

Cassell, J. (2007). *Body language : Lessons from the near-human*. Chicago : University of Chicago Press.

Chen, Y., Gao, W., Zhu, T., and Ling, C. (2002). Learning prosodic patterns for mandarin speech synthesis. *Journal of Intelligent Information Systems*.

Chen, Y.-L., Chiang, M.-C., and Ko, M.-T. (2003). Discovering time-interval sequential patterns in sequence databases. *Journal of Expert Systems with Applications*.

Chindamo, M., Allwood, J., and Ahlsen, E. (2012). Some suggestions for the study of stance in communication. In *Proceedings of the International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation : Encoder-decoder approaches. *arXiv preprint arXiv :1409.1259*.

Chollet, M., Ochs, M., and Pelachaud, C. (2013). A multimodal corpus for the study of non-verbal behavior expressing interpersonal stances. In *Proceedings of the Workshop Multimodal Corpora : Beyond Audio and Video, hosted by the International Conference on Intelligent Virtual Agents*.

Chollet, M., Ochs, M., and Pelachaud, C. (2014). From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression. In *Proceedings of the International Conference on Intelligent Virtual Agents*.

Clavel, C., Vasilescu, I., Devillers, L., Richard, G., and Ehrette, T. (2008). Fear-type emotion recognition for future audio-based surveillance systems. *Journal of Speech Communication*.

Cowie, R., Gunes, H., McKeown, G., Vaclau-Schneider, L., Armstrong, J., and Douglas-Cowie, E. (2010). The emotional and communicative significance of head nods and

shakes in a naturalistic database. In *Procedings of LREC International Workshop on Emotion*.

Csurka, G. (2017). Domain adaptation for visual applications : A comprehensive survey. *arXiv preprint arXiv :1702.05374*.

De Jong, N. H. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Journal of Behavior research methods*.

Dechter, R., Meiri, I., and Pearl, J. (1991). Temporal constraint networks. *Journal of Artificial intelligence*.

Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. (2014). Covarep - a collaborative voice analysis repository for speech technologies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Dermouche, S. and Pelachaud, C. (2016). Sequence-based multimodal behavior modeling for social agents. In *Proceedings of the International Conference on Multimodal Interaction*.

D'Errico, F., Signorello, R., Demolin, D., and Poggi, I. (2013). The perception of charisma from voice : A cross-cultural study. In *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*.

Deutsch, F. M., LeBaron, D., and Fryer, M. M. (1987). What is in a smile ? *Journal of Psychology of Women Quarterly*.

Ding, C., Xie, L., and Zhu, P. (2015a). Head motion synthesis from speech using deep neural networks. *Journal of Multimedia Tools and Applications*.

Ding, C., Zhu, P., and Xie, L. (2015b). Blstm neural networks for speech driven head motion synthesis. In *Proceedings of the Conference of the International Speech Communication Association*.

Dousson, C. and Duong, T. V. (1999). Discovering chronicles with numerical time constraints from alarm logs for monitoring dynamic systems. In *Proceedings of the IJCAI*.

Du Bois, J. W. (2007). The stance triangle. *Journal of Stancetaking in discourse : Subjectivity, evaluation, interaction*.

Ekman, P. and Friesen, W. (1978). Facial action coding system : a technique for the measurement of facial movement. *Journal of Palo Alto : Consulting Psychologists*.

Freedman, M. B., Leary, T. F., Ossorio, A. G., and Goffey, H. S. (1951). The interpersonal dimension of personality. *Journal of personality*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*.

Goleman, D. (2006). The socially intelligent. *Journal of Educational leadership*.

Greenwood, D., Laycock, S., and Matthews, I. (2017). Predicting head pose from speech with a conditional variational autoencoder. In *Interspeech*.

Guillame-Bert, M. and Crowley, J. L. (2012). Learning temporal association rules on symbolic time sequences. In *Proceedings of the Asian Conference on Machine Learning*.

Gurtman, M. B. (2009). Exploring personality with the interpersonal circumplex. *Journal of Social and Personality Psychology Compass*.

Haag, K. and Shimodaira, H. (2016). Bidirectional lstm networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In *Proceedings of the International Conference on Intelligent Virtual Agents*.

Hall, J. A., Coats, E. J., and LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations : a meta-analysis. *Psychological bulletin*.

Henley, N. M. (1995). Body politics revisited : What do we know today. *Journal of Gender, power, and communication in human relationships*.

Hess, U., Beaupré, M. G., Cheung, N., et al. (2002). Who to whom and why–cultural differences and similarities in the function of smiles. *Journal of An empirical reflection on the smile*.

Hess, U., Blairy, S., and Kleck, R. E. (2000). The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal behavior*.

Hess, U. and Thibault, P. (2009). Why the same expression may not mean the same when shown on different faces or seen by different people. *Journal of Affective information processing*.

Hirate, Y. and Yamana, H. (2006). Generalized sequential pattern mining with item intervals. *JCP*.

Ho, T. B., Nguyen, T. D., Kawasaki, S., Le, S. Q., Nguyen, D. D., Yokoi, H., and Takabayashi, K. (2003). Mining hepatitis data with temporal abstraction. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Journal of Neural computation*.

Horowitz, L. M., Wilson, K. R., Turan, B., Zolotsev, P., Constantino, M. J., and Henderson, L. (2006). How interpersonal motives clarify the meaning of interpersonal behavior : A revised circumplex model. *Journal of Personality and Social Psychology Review*.

Janssoone, T. (2015). Temporal association rules for modelling multimodal social signals. In *proceedings of the International Conference on Multimodal Interaction (doctoral consortium)*.

Janssoone, T., Clavel, C., Bailly, K., and Richard, G. (2016a). Des signaux sociaux aux attitudes : de l´utilisation des règles d´association temporelle. In *proceedings of the WACAI 2016, Workshop . Affect . Compagnon Artificiel . Interaction*.

Janssoone, T., Clavel, C., Bailly, K., and Richard, G. (2016b). Using temporal association rules for the synthesis of embodied conversational agents with a specific stance. In *proceedings of the International Conference on Intelligent Virtual Agents*.

Janssoone, T., Clavel, C., Bailly, K., and Richard, G. (2017). Règles d'associations temporelles de signaux sociaux pour la synthèse de comportements d´agents conversationnels animés : application aux attitudes sociales. *Revue d'Intelligence Artificielle*.

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

Kiesler, D. J. (1996). From communications to interpersonal theory : A personal odyssey. *Journal of personality assessment*.

Knapp, M., Hall, J., and Horgan, T. (2013). *Nonverbal communication in human interaction*. Cengage Learning.

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. (2017). Afew-va database for valence and arousal estimation in-the-wild. *Journal of Image and Vision Computing*.

Lan, X., Li, X., Ning, Y., Wu, Z., Meng, H., Jia, J., and Cai, L. (2016). Low level descriptors based dblstm bottleneck feature for speech driven talking avatar. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*.

Laskowski, K., Edlund, J., and Heldner, M. (2008). Learning prosodic sequences using the fundamental frequency variation spectrum. In *Procedings of the 4th International Conference on Speech Prosody*.

Leary, T. (1958). Interpersonal diagnosis of personality. *American Journal of Physical Medicine & Rehabilitation*.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Journal of Neural computation*.

Lee, J. and Marsella, S. (2012). Modeling speaker behavior : A comparison of two approaches. In *Proceedings of the Intelligent Virtual Agents*.

Li, L., McCann, J., Pollard, N. S., and Faloutsos, C. (2009). Dynammo : Mining and summarization of coevolving sequences with missing values. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Li, L., Yang, D.-Z., and Shen, F.-C. (2010). A novel rule-based intrusion detection system using data mining. In *Proceedings of the IEEE International Conference on Computer Science and Information Technology,*.

Liang, Z., Xinming, T., Lin, L., and Wenliang, J. (2005). Temporal association rule mining based on t-apriori algorithm and its typical application. In *Proceedings of the International Symposium on Spatial-Temporal Modeling Analysis*.

Locke, K. D. (2006). Interpersonal circumplex measures. *Journal of S. Strack (Ed.), Differentiating normal and abnormal personality*.

Mannila, H. and Toivonen, H. (1996). Discovering generalized episodes using minimal occurrences. In *Proceedings of the KDD*.

Mannila, H., Toivonen, H., and Verkamo, A. I. (1997). Discovery of frequent episodes in event sequences. *Journal of Data mining and knowledge discovery*.

Marsella, S., Gratch, J., and Petta, P. (2010). Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*.

Martínez, H. P. and Yannakakis, G. N. (2011). Mining multimodal sequential patterns : a case study on affect detection. In *Proceedings of the 13th International Conference on Multimodal Interfaces*.

Masseglia, F., Poncelet, P., and Teisseire, M. (2003). Incremental mining of sequential patterns in large databases. *Journal of Data & Knowledge Engineering*.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and 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*.

Mertens, P. (2004). The prosogram : Semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of the Speech Prosody International Conference*.

Mignault, A. and Chaudhuri, A. (2003). The many faces of a neutral face : Head tilt and perception of dominance and emotion. *Journal of Nonverbal Behavior*.

Morency, L.-P., de Kok, I., and Gratch, J. (2008). Predicting listener backchannels : A probabilistic multimodal approach. In *Proceedings of the Intelligent Virtual Agents*.

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*.

Motulsky, H. (2013). *Intuitive biostatistics : a nonmathematical guide to statistical thinking*. Oxford University Press, USA.

Murphy, N. A. (2005). Using thin slices for behavioral coding. *Journal of Nonverbal Behavior*.

Naqvi, M., Hussain, K., Asghar, S., and Fong, S. (2011). *Mining Temporal Association Rules with Incremental Standing for Segment Progressive Filter*.

Nicolle, J., Bailly, K., and Chetouani, M. (2016). Real-time facial action unit intensity prediction with regularized metric learning.

Ochs, M. and Pelachaud, C. (2012). Model of the perception of smiling virtual character. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*.

Ochs, M. and Pelachaud, C. (2013). Socially aware virtual characters : the social signal of smiles [social sciences]. *IEEE Signal Processing Magazine*.

Ochs, M., Prepin, K., and Pelachaud, C. (2013). From emotions to interpersonal stances : Multi-level analysis of smiling virtual characters. In *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*.

Pappas, N. and Popescu-Belis, A. (2013). Combining content with user preferences for ted lecture recommendation. In *Proceedings of the 11th International Workshop on Content Based Multimedia Indexing*.

Pecune, F., Cafaro, A., Chollet, M., Philippe, P., and Pelachaud, C. (2014). Suggestions for extending saiba with the vib platform. In *Proceedings of the Workshop on architectures and standards for IVAs, held at the '14th international conference on intelligent virtual agents'*.

Pecune, F., Cafaro, A., Ochs, M., and Pelachaud, C. (2016a). Evaluating social attitudes of a virtual tutor. In *Proceedings of the International Conference on Intelligent Virtual Agents*.

Pecune, F., Ochs, M., Marsella, S., and Pelachaud, C. (2016b). Socrates : from social relation to attitude expressions. In *Proceedings of the International Conference on Autonomous Agents & Multiagent Systems*.

Pentland, A. (2004). Social dynamics : Signals and behavior. In *Proceedings of the 3rd International Conference on Developmental Learning*.

Piolat, A. and Bannour, R. (2008). Emotions et affects : contribution de la psychologie cognitive. *Journal of Le sujet des émotions au Moyen Age*.

Ravenet, B., Ochs, M., and Pelachaud, C. (2013). From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In *Proceedings of the International Workshop on Intelligent Virtual Agents*.

Rehm, M. and André, E. (2008). From annotated multimodal corpora to simulated human-like behaviors. *Journal of Modeling Communication with Robots and Virtual Humans*.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*.

Robinson, M. D., Zabelina, D. L., Ode, S., and Moeller, S. K. (2008). The vertical nature of dominance-submission : Individual differences in vertical attention. *Journal of Research in Personality*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Sacchi, L., Larizza, C., Combi, C., and Bellazzi, R. (2007). Data mining with temporal abstractions : learning rules from time series. *Journal of Data Mining and Knowledge Discovery*.

Sadoughi, N., Liu, Y., and Busso, C. (2017). Meaningful head movements driven by emotional synthetic speech. *Journal of Speech Communication*.

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

Snoek, C. G., Worring, M., and Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*.

Srikant, R. and Agrawal, R. (1996). Mining sequential patterns : Generalizations and performance improvements. *Journal of Advances in Database Technology*.

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama : learning lip sync from audio. *ACM Transactions on Graphics (TOG)*.

Tiedens, L. Z. (2000). Powerful emotions : The vicious cycle of social status positions and emotions. *Journal of Emotions in the workplace : Research, theory, and practice*.

Trapnell, P. D. and Broughton, R. H. (2006). The interpersonal questionnaire (ipq) : Duodecant markers of wiggins' interpersonal circumplex.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features ? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Truong, K., Heylen, D., Chetouani, M., Mutlu, B., and Salah, A. A. (2015). the international workshop on emotion representations and modelling for companion technologies. In *Proceedings of International Conference on Multimodal Interaction (Workshop)*.

Tusing, K. J. and Dillard, J. P. (2000). The sounds of dominance. *Journal of Human Communication Research*.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). Avec 2016 : Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*.

Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. (2009a). Canal9 : A database of political debates for analysis of social interactions. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*.

Vinciarelli, A. and Mohammadi, G. (2014). A survey of personality computing. *IEEE Transactions on Affective Computing*.

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009b). Social signal processing : Survey of an emerging domain. *Journal of Image and vision computing*.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., and Schröder, M. (2012). Bridging the gap between social animal and unsocial machine : A survey of social signal processing. *IEEE Transactions on Affective Computing*.

Wallbott, H. G. and Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of personality and social psychology*.

Ward, N. G. and Abu, S. (2016). Action-coordinating prosody. In *Proceedings of the Speech Prosody*.

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms : The interpersonal domain. *Journal of personality and social psychology*.

Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Xu, L., Zhiyong, W., Helen, M. M., Jia, J., Xiaoyan, L., and Lianhong, C. (2016). Expressive speech driven talking avatar synthesis with dblstm using limited amount of emotional bimodal data. In *Proceedings of the INTERSPEECH*.

Youssef, A. B., Chollet, M., Jones, H., Sabouret, N., Pelachaud, C., and Ochs, M. (2015). Towards a socially adaptive virtual agent. In *Proceedings of the International Conference on Intelligent Virtual Agents*.

Yu, H., Gui, L., Madaio, M., Ogan, A., Cassell, J., and Morency, L.-P. (2017). Temporally selective attention model for social and affective state recognition in multimedia content. *Journal of Multimedia Content*.

Zhao, R., Sinha, T., Black, A., and Cassell, J. (2016). Socially-aware virtual agents : Automatically assessing dyadic rapport from temporal patterns of behavior. In *Proceedings of the International Conference on Intelligent Virtual Agents*.