SMART LABEX
SMART HUMANMACHINEHUMAN INTERACTIONS IN THE DIGITAL

Rapport de Mi-Parcours de Thèse

Project SENSE

Encadrants : Frederic Bevilacqua, Catherine Pelachaud
Participants : Labex SMART, IRCAM, LTCI-Telecom-Paristech

October 11, 2016

http://www.smart-labex.fr/

Laboratory of Excellence SMART (ANR-11-LABX-65) is supported by French State funds managed by the ANR within the Investissements d'Avenir Programme under reference ANR-11-IDEX-0004-02

Contents

I Introduction

Au cour du projet de thèse "Adaptation au cours du temps de l'interaction" du projet Sense, notre objectif est de développer un système adaptatif de gestion de l'interaction. Nous considérons ici l'interaction entre des personnes et/ou des agents virtuels, caractérisée par un ensemble de comportements, d'actions et de réponses réciproques ainsi que des aspects de communication verbales et non verbales. En particulier, nous nous proposons de considérer l'interaction comme un processus émergent [Simpson and Galbo, 1986].

Afin de rendre compte de cette dimension émergente de l'interaction, le modèle régissant (ou décrivant l'interaction) doit de fait être adaptatif sur plusieurs aspects. Ne pouvant adresser l'ensemble des phénomènes d'adaptation, nous avons décidé de nous concentrer sur la dimension temporelle de l'adaptation. Par exemple, les phénomènes de synchronisation peuvent être considérés comme des cas d'adaptations temporelles de l'interaction [Chetouani, 2014].

La dimension multimodale de l'interaction est également un point important que nous considérons, l'interaction n'ayant pas lieu uniquement au niveau du langage mais également dans toutes les dimensions non-verbales de l'interaction [Mehrabian, 1977]. En particulier, nous étudierons comment la communication non-verbale intervient dans la régulation de l'interaction et plus particulièrement d ans la gestion des tour de parole (turn-taking) [Thórisson, 2002].

Nous nous intéresserons plus particulièrement à deux cas d'applications : les Agents Conversationnels Animés (ACA, ou Embodied Conversational Agents, ECA) [Cassell et al., 1999] et des Agents Improvisateurs Créatifs [Assayag and Dubnov, 2004]. Nous faisons l'hypothèse que le développement d'un modèle qui pourra s'appliquer à ces deux cas d'applications sera suffisamment générique pour général pour pouvoir s'appliquer à d'autres cas d'applications.

II Contexte

Notre recherche s'intéresse à la régulation de la parole dans la conversation et à l'Interaction Homme-Machine, ainsi que pour la partie applicative aux domaines des Agents Conversationnels Animés (ACA) et des Agents Musicaux Créatifs (AMC). La richesse de l'interaction étant essentielle pour l'engagement d'utilisateurs humains [Novielli et al., 2010], et pour enrichir cette interaction nous proposons l'usage des signaux non-verbaux, sachant que la synchronie non-verbale peut aider à la construction de relations sociales entre interactants [LaFrance, 1982].

Le comportement non-verbal peut être décrit comme " toutes les actions distinctes de la paroles" [Mehrabian, 1977], bien que certains aspects paralinguistiques comme la prosodie y soient également inclus. La communication non-verbale peut prendre différentes formes et s'exprimer de diverses manières (appelées modalités), comme le regard citekendon1967some ou des signaux paralinguistiques [Goodwin, 1981, Allwood, 1976]. D'après Knapp et al., tous les humains sont naturellement experts dans la communication multimodale [Knapp, 2012], ce qui signifie qu'ils sont capables de recevoir et d'émettre simultanément des signaux sur différentes modalités, qu'ils en aient consicence ou non. D'après Argyle, la communication non verbale remplit quatre fonctions principales : exprimer les émptions, transmettre des attitudes interpersonnelles, présenter sa personalité et accompagner la parole. cette dernière fonction est essentielle dans la régulation de la conversation.

Un tour dans la conversation peut être définie comme le moment entre la prise de parole et et son abandon, que ce dernier puisse être consensuel ou forcé [Goodwin, 1981]. L'expression de la parole n'est pas évidente et peut prendre plusieurs sens : des tours se chevauchant peuvent indiquer un conflit entre locuteurs mais égelement un haut niveau de synchronie puisque les interactants sont capables de déchiffrer les indices de l'abandon du tour [Allwood, 1995]. Les mécanismes de régulation de la conversation sont perceptibles à travers les signaux émis simultanément par les locuteurs et les auditeurs. Duncan identifie trois types de signaux de régulation : des signaux de passage de la parole, des signaux pour conserver ou prendre la parole face à d'autres locuteurs et des backchannels, qui peuvent exprimer de multiples attentions, de la simple reconnaissance à jusqu'à l'expression de l'état mental de l'émetteur [Allwood, 1976]. Ces signaux sont essentiels à la conversation et garantissent sa fluidité : Ten Bosch et al. ont par exemple montré que dans des conversations téléphoniques les pauses entre les temps de parole des interlocuteurs seraient 20 à 50% plus long que dans les conversations en face à face et le nombre de chevauchement entre deux teps de parole augmente de 70 %, ce qui indique que ces phases de transitions doivent être prises en compte dans un modèle décrivant la conversation.

Le modèle de Sacks [Sacks et al., 1974] émet l'hypothèse que comme dans tous les contextes organisés (débats politiques, jeux de plateau, carrefours routiers, etc.), des tours surviennent naturellement dans la conversation. Sacks divise la conversation en tours dans lesquels existent des endroits où la transition est pertinente, qui surviennent en fonction du contexte interactif (par exemple, quand le locuteur indique qu'il est sur le point de s'arrêter de parler et indique son choix pour l'interlocutuer suivant). En se basant sur un corpus d'enregistrement et leurs transcritpions, Sacks définit un ensemble de règles qu'il pense réguler la conversation. Trois types de règles émergent de son travail :

1. Un tour est décrit comme une entitée ductile et organisée, où les locuteurs parlent la plupart du temps.

2. Les tours dans la conversation sont des entités dynamiques dont la longueur n'est pas prédéterminée.

3. Il existe des techniques pour réguler le tour conversationnel.

De ses règles, une constante se révèle : les tours de paroles sont certes organisées par des règles communes, cependant leur expression est dynamique et émerge de ces règles.

Clark [Clark, 1996] va encore plus loin dans cette affirmation : d'après lui, les règles que Sacks a découvertes sont des régularités que Sacks a trouvé dans l'interaction mais sont incapbles de décrire pleinement l'interaction (par exemple, le fait que certaines personnes parlent les unes après les autres en tentant de minimiser le chevauchement et les pauses entre les temps de parole est présent dans certaines cultures mais pas dans toutes [Kilpatrick, 1986]). Clark affirme que la notion de tour est en elle-même émergente et prend racine dans les comportements d'écoute et de parole. Il ajoute que que les règles de changement de tour pouvaient être plus flexibles que définies par les modèles historiques tels que celui de Sacks [Clark and Krych, 2004]. EN accord avec la description de Clark, nous supposons qu'un model de conversation devra prendre en compte les comportements d'écoute et de parole.

III Travaux Connexes

Les agents virtuels (qu'il s'agisse d'ACA ou de robots) ont de plus en plus adressées les questions de recherches liées à la répratition de la parole ces dernières années. Une des premières questions est la modélisation, et si nous adresserons plus précisément cette question dans la section IV, nous pouvons déjà affirmer qu'une des premières manières de décrire les comportements identifiés par Sacks et Clark était de faire usage d'un Automate à Etats Finis (AEF) (comme dans [Raux and Eskenazi, 2009], [Thórisson, 2002] ou [Ravenet et al., 2015]). Plus récemment, une nouvelle manière de décrire la répratition de la parole a vu le jour : l'emploi d'une variable simple (que pour résumer on pourrait présenter comme booléenne) qui a chaque instant indique au système si l'agent doit parler ou non à l'instant t (comme dans [Kose-Bagci et al., 2008] ou [Bohus and Horvitz, 2011]).

Dans les deux cas, le système alterne entre des états qui décrivent son comportement conversationnel, et la manière dont ces transitions ont lieu est déterminée par des équations mathématiques plus ou moins complexes dont les variables sont liées aux données de l'interaction. Ces données peuvent être de nature unimodale (comme le fait que l'agent parle ou non dans [Raux and Eskenazi, 2009] ou la longueur du tour précédent comme dans [Kose-Bagci et al., 2008]), mais elles peuvent aussi en plus de ces simples dimensions prendre en compte les multiples modalités de l'interaction non verbale, comme le mouvement [Bohus and Horvitz, 2011], le regard [Chao and Thomaz, 2012] ou la prosodie [Jonsdottir and Thórisson,].

Les équations en elle-même peuvent varier en complexité. Dans Kose-Bagçi et al. [Kose-Bagci et al., 2008] où un humain et un robot interacte en frappant alternativement sur un tambour avec l'objectif de minimiser les pauses et les chevauchements entre les temps de paroles, cette equation est volontairement simple ; des formules comme $\frac{x}{seuil} = 1$ déterminent si l'agent doit continuer à) frapper sur le tambour ou non, où x est le nombre de pulsaitons frappées sur le tambour et $seuil$ est le nombre de pulsation que l'interlocuteur humain a frappées au tour précédent. Le but de telles équations est de faire émerger l'interaction entre les participants en attendant de l'humain qu'il reconnaisse les motifs dans le tour de l'autre participant. A l'inverse, dans le système de Ravenet et al. [Ravenet et al., 2015] le modèle décrivant les états conversationnels est plus complexe (un automate à état fini), et les transitions d'un état à l'autre changent selon l'éta courant, et prennent en compte des données de l'interaction comme la dominance et l'amicalité d'un agent par rapport à un autre. En définissant plus de règles, le système est certes moins souple mais cherche à être plus précis dans sa descritption de l'interaction conversationnelle.

Certains systèmes prennent directement en compte le temps en le corrélant avec les données observés. Dans le système de Chao et al. [Chao and Thomaz, 2012], l'agent observe le regard, la quantité de mouvement et les motifs réuccrents dans la parole dans un jeu de "Jaquadit" et incorpore les résultats dans un Réseau de Petri Temporalisé. Les occurences de certains évènements à une temporalité donnée (qui a été déterminée par des observations d'interactions réelles) est ce qui change le comportement et déclenche sa décision de parler ou non.

IV Description du système

La manière dont nous avons choisi de modéliser notre comportement conversationnele est un AEF. Cette modélisation est déjà présente dans la littérature et se justifie en accord avec Gibson, qui indique que les comportements conversationnels reflètent les états mentaux ddes locuteurs [Gibson, 2003]. Deux types de modélisation sont utilisées : soit l'AEF modélise l'état global de la conversation (comme par exemple [Raux and Eskenazi, 2009], où l'agent a vue sur le système entier et si la parole est prise par lui-même ou un autre agent) ou l'AEF modélise les états internes de l'agent (comme dans [Jonsdottir and Thórisson,] où l'AEF décrit si l'agent veut parler, est en train de parler ou écoute).

Modéliser la conversation dans son entier dans un AFE se rapprocherait du modèle de coversation de Sacks ; en modélisant le système, les concepteurs de tels agents décident à l'avance des *règles conversationnelles*, alors qu'un AEF modélisant les états de l'agent décrit les *règles comportementales* dont les *règles conversationnelles* doivent émerger. Cette dernière hypothèse suit la thèse de Goffman, qui suppose qu'il existe une relation fonctionnelle entre la stucruture du moi et la structure de l'interaction [Goffman, 1955]. Ces modèles s'appuient sur la décision ou non de parler, comme c'est le cas du modèle de Bohus et Horowitz [Bohus and Horvitz, 2011], où seule la décision de prendre la parole est modélisée et pas les comportements non verbaux afférents. En nous inspirant de ces modèles, nous pensons que les états de "Parole" et "d'Ecoute" doivent être inclus au modèle, et qu'ils doivent pas se limiter à la prise de parole mais dovent être liés aux comportements non-verbaux.

Cependant, d'après Sacks, le nombre de participants peut varier au cours de la conversation [Sacks et al., 1974] ; d'où l'adjonction d'un état "Au Repos" qui représentera l'état conversationnel interne de l'agent quand il n'écoute et ne parle plus mais reste dans le même espace physique que les agents qui continuent à converser. Cet état doit être également l'état initial, car au début de l'interaction personne ne parle où n'écoute.

Enfin, les transitions entre les temps de parole, qu'il s'agisse de l'allocation ou la libération du tour, font partie intégrante de la conversation puisqu'ils permettent une bonne coordination entre les tours [Ten Bosch et al., 2005] (même lorssqu'ils s'affrontent pour la possession de la parole [Bruijnes et al., 2012]). Il semble nécessaire d'ajouter à notre modèles des états décrivant le comportement de demande de la parole et de libération de la parole. Notre modèle est représenté dans la figure 1.

V Implémentation

L'interaction conversationnelle a de spéciale qu'elle laisse une grande part à l'inférence ; d'après Grice [Grice, 1991], nous infèrerions le comportement conversationnel de nos interlocuteurs en fonctions des croyances que nous aurions sur eux. En accord avec cette théorie, nous avons décidé d'employer des Modèles de Markovs Cachés (Hidden Markov Model, HMM [Baum and Petrie, 1966]) pour implémenter le modèle mentionné dans la section IV.

Les HMMs sont des modèles probabilistes respectant la propriété de Markov, i.e. le fait que l'état courant est déterminé uniquement en fonction de l'état précédent. La différence avec d'autres modèlmes probabilistes est que l'état courant

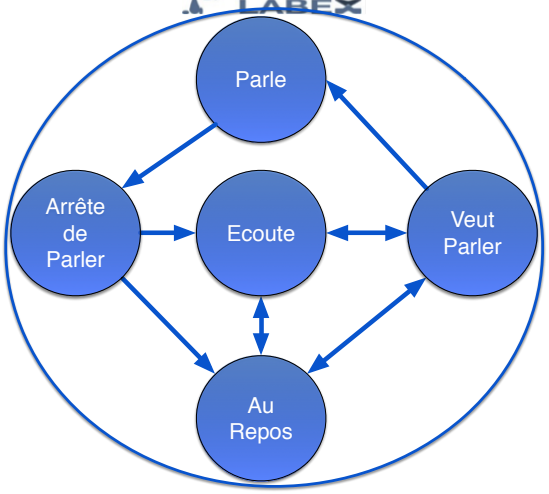

Figure 1: The Finite State Machine modeling the conversational behavior of our agent

d'un HMM n'est jamais certain mais est seulement le plus probable à l'instant t en fonction des observations à cet instant. Ainsi, le HMM reflète l'idée de Grice de l'inférence de l'état conversationnel à partir des connaissances et de ce qui est visible extérieurement.

Un des autres avantages des HMMs est que le lien entre Etats Cachés et Emissions est lui aussi probabiliste; ainsi et contrairement à un simple AEF, le système est non déterministe : les mêmes causes (parole, émission d'une modalité non-verbale, etc.) peuvent engendrer des effets différents (transition d'un état A à un état B, transition d'un état A à un état C, conservation de l'état courant, etc.).

Un troisième avantage des HMMs est que les liens probabilistes entre les différents états et les liens probabilistes entre états et émissions sont appris. De cette manière, notre système devient adaptable : en fonction des données apprises, l'agent aura un comportement différent, ce qui lui permet d'être employé dans divers cas applicatifs, comme dans le cas de notre projet de thèse où l'agent peut être un interlocuteur dans une conversation ou un agent musical qui doivent adapoter des comportements différents.

Cependant, lors de la conversation, nous n'inférons pas seulement notre propre comportement mais également celui des autres interlocuteurs et notre comportement évolue en fonction de celui que nous avons inféré chez nos interlocuteurs. Afin de modéliser ce lien et de conserver les propriétés intéressantes des HMM. Pour répondre à ces deux problèmmes, nous avons fait le choix d'employer des modèles d'influence [Dong et al., 2007].

Un modèle d'influence comprend plusieurs HMMs dont les probabilités de transition dépendent non seulement de l'état présent du HMM mais également de l'état présent des autres HMMs selon une pondération appelée influence qui peut elle aussi être apprise à partir de données réelles. Un autre avantage des modèles d'influences est qu'ils permettent de simplifier l'apprentissage des rapports d'un modèle à l'autre, ce qui ne serait pas nécessairement le cas avec des HMMs couplés.

V.1 Formalisation

La représentation graphique de notre modèle d'influence est présenté dans la figure 2. En représentant le système par un modèle d'influence et chaque agent par des HMMs de ce modèle, nous voulons non seulement représenter les comportements des agents mais également les influences qu'ils peuvent exercer les uns sur les autres.

Chaque agent contient les états cachés S_iAR, VP, P, ADP,E, où

- AR = Au Repos, l'état dans lequel l'agent n'émet aucun signal mais est présent dans le même espace géographique que d'autres agents
- VP = Veut Parler, l'état dans lequel l'agent signale aux autres agents qu'il désire prendre la parole
- P = Parole, l'état dans lequel l'agent s'adresse à un ou plusieurs autres agents
- ADP = Arrête De Parler, l'état dans lequel l'agent signale aux autres agents qu'il a fini de s'adresser à eux
- E = Ecoute, l'état dans lequel l'agent signale à un ou plusieurs locuteurs qu'il les écoute

Notons la probabilité P_i d'être dans l'état S_i. Cette probabilité est initialisée au début du processus à (1, 0, 0, 0, 0) et la matrice de transtion est notée A_ij. Cette matrice est apprise par les agents à partir des données fournies au modèle (dans le cas présent des séquences d'interactions générées grâce à l'interface utilisée dans Ravenet et al. [Ravenet et al., 2015]). Pour générer ces données, nous faisons parler entre eux quatre agents VIB [Pecune et al., 2014] et enregistront trois données tous les quarts de seconde : l'état mental, si l'agent parle et si l'agent fait ou non un geste.

A chaque étape t, une observation de l'un des autres agents est effectuée et décrite par le vecteur
$O(t)$ =(Sound, Gesture) = (0,1)
où, par exemple, Sound = 0 signifie que l'agent observé n'émet aucun son à l'instant t et Gesture = 1 signifie que l'agent observé a effectué un geste à l'instant t. Ces données sont discrètes (on enregistre ou non la présence d'un des descripteurs). Ces descripteurs sont pour le moment binaires, mais il est tout à fait envisageable d'employer des descritpeurs plus complexes (que ce soit au niveau de l'augmentation des valeurs numériques possibles ou l'emploi de descripteurs composés).

Nous pouvons alors calculer la probabilité de l'agent c d'être dans l'état s

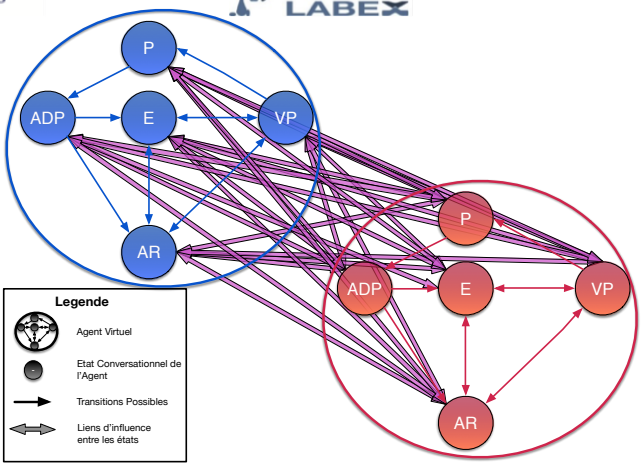

Figure 2: Représentation du modèle d'influence

$(P_(S^c_{t+1} = s))$

$$P_(S^c_{t+1} = s) = \sum_{c_1=1}^{C} \sum_{s_1=1}^{m} c1 h^{(s1,s)}_{(c_1,c)} P_a(S_t = s_1)$$

$$P_(S^c_{t+1} = s) = \sum_{c_1=1}^{C} (h^{(AR,s)}_{(c_1,s)} P_(S^c_t = AR) + h^{(VP,s)}_{(c_1,s)} P_(S^c_t = VP) + h^{(P,s)}_{(c_1,s)} P_(S^c_t = P) + h^{(AP,s)}_{(c_1,s)} P_(S^c_t = AP) + h^{(E,s)}_{(c_1,s)} P_(S^c_t = E))$$

où :

- $\{1, 2, ..., C\}$ sont les différents agents du système
- $\{s_1, s_2, ..., s\} = \{AR, VP, P, AP, E\}$ sont les états possibles pour chaque agent
- $h^{(s_1,s)}_{(c_1,c)}$ est la valeur d'influence de l'agent c_1 sur l'agent c pour l'état de c s et l'état de c_1 s_1. $h^{(s_1,s)}_{(c_1,c)} = d^{(c_1,c)} a^{(c_1,c)}_{s1,s}$ où $d^{(c_1,c)}$ représente l'influence de l'agent c_1 sur l'agent c et $a^{(c_1,c)}_{s1,s}$ représente l'infuence de l'état s_1 de l'agent c_1 sur l'état s de l'agent c. Ces grandeurs sont apprises en même temps que l'apprentissage des matrices de transitions par un algorithme semblable à celui de Viterbi (pour plus d'informations, se référer à [Dong et al., 2007]). De ce fait, l'apprentissage des données ne se fera pas en temps réel mais devra être fait en amont de l'utilisation du modèle.

V.2 Résultats préliminaires

Pour le moment, nous avons commencé par récolter des données interactives en enregistrant des conversations utilisant l'interface de l'article de Ravenet et al [Ravenet et al., 2015] où le modèle de conversation de la section IV a été implémenté. Ces données comprenaient trois types d'observations : si l'agent était en train de parler, si l'agent bougeait ses bras et si l'agent avait son visage en mouvement. Le système disposait d'une horloge interne qui réintérrogeait l'état de chaque agent tous les quarts de seconde. Afin de modifier les temps de parole et ce à qui ils allaient s'adresser, nous avons fait varier les valeurs d'amiclaité et de dominance entre les agents.

Ces données ont ensuite été entrées dans la toolbox des modèles d'influence de matlab [1] créée par l'équipe à l'origine de l'utilisation de ces modèles d'influence dans un cadre d'interaction homme-machine [Dong et al., 2007]. Nous avons utilisées cette toolbox pour générer des données de même nature (i.e. multimodales) que celles ayant servi à l'apprentissage. Nous avons constaté que lors de la phase d'apprentissage les matrices de transition de chaque agent convergeaient, montrant qu'après l'apprentissage le modèle restait stable. Ceci est visible dans la figure 3, qui a été obtenue en moyennant les valeurs de la matrice de transition d'un agent sur 500 itérations de l'algorithme d'apprentissage.

Lorsque nous avons essayé de générer des séquences avec ces modèles d'influence, nous avons fait deux observations : premièrement, les temps de paroles sont significativement les mêmes entre les agents initiaux et ceux générés par les modèles d'influence. Enfin, les comportements observés sont identiques à ceux des agents intiiaux : si par exemple un des agents monopolisait le temps de parole et les autres le laissaient faire, le modèle d'influence générait une interaction qui reflétait ce comportement. Ceci est présenté dans le tableau ci-dessous, qui montre les temps générés par le modèle Unity ($\{A_{init}, B_{init}, C_{init}, D_{init}\}$), et les temps de parole moyennés sur 500 itérations de chaque agent générés par les modèles d'influence ($\{A_{gen}, B_{gen}, C_{gen}, D_{gen}\}$).

Temps de Parole (secondes)								
Agents	A_{init}	B_{init}	C_{init}	D_{init}	A_{gen}	B_{gen}	C_{gen}	D_{gen}
Temps de parole égaux	45.2	43.1	43.7	46.2	48.1	42.3	45.2	43.3
Un agent monopolise la parole	84.2	30.1	20.2	39.4	78.1	28.1	24.2	34.6
Les agents parlent en même temps	85.8	78.7	79.4	77.3	82.2	80.1	82.0	79.9

VI Travaux futurs

La suite de notre travail de thèse suivra plusieurs phases. D'abord, nous avons l'intention d'intégrer les modèles d'influences à l'intérieur de l'interface GRETA-UNITY employée dans l'article de Ravenet et al. [Ravenet et al., 2015]. Une fois cette intégration réalisée, nous utiliserons ce système pour réaliser deux études : une quantitative dans environ un mois et une qualitative dans deux mois. Dans l'étude quantitative, nous avons l'attention d'utiliser les outils de mesure de la synchronie de Varni et al [Varni et al., 2015] afin d'évaluer l'émergence de synchronie entre les agents et voir si notre modèle parvient à une meilleur synchronie entre les agents conversationnels que celui de Ravenet et al. et les HMMs simples.

[1] http://vismod.media.mit.edu/vismod/demos/influence-model/software-usage.htm

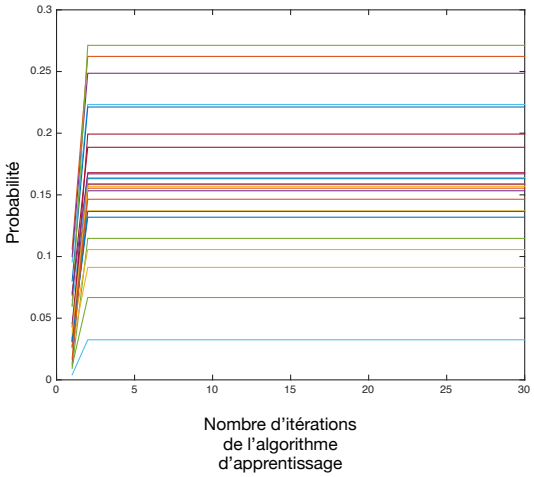

Figure 3: Evolution des valeurs des matrices de transition pour un agent au cours de l'apprentissage

Nous effecturerons ensuite une étude qualitative afin de faire évaluer par des sujets naïfs la qualité de l'interaction, sa vraisemblance et si les comportements observés sont bien ceux que nous espérions faire émerger de l'interaction des modèles individuels.

Nous comptons ensuite intégrer notre système au système OMaX [Lévy et al., 2012], qui aujourd'hui encore a besoin d'un opérateur humain pour signifier au système quand il doit jouer ou non dans trois à cinq mois, avant de rédiger la thèse au alentours de février 2017.

Bibliography

[Allwood, 1976] Allwood, J. (1976). *Linguistic communication as action and cooperation*. PhD thesis, Gothenburg University.

[Allwood, 1995] Allwood, J. (1995). An activity based approach to pragmatics.

[Assayag and Dubnov, 2004] Assayag, G. and Dubnov, S. (2004). Using factor oracles for machine improvisation. *Soft Computing*, 8(9):604–610.

[Baum and Petrie, 1966] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563.

[Bohus and Horvitz, 2011] Bohus, D. and Horvitz, E. (2011). Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*, pages 98–109. Association for Computational Linguistics.

[Bruijnes et al., 2012] Bruijnes, M. et al. (2012). Computational models of social and emotional turn-taking for embodied conversational agents: a review.

[Cassell et al., 1999] Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., and Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 520–527. ACM.

[Chao and Thomaz, 2012] Chao, C. and Thomaz, A. L. (2012). Timing in multimodal turn-taking interactions: Control and analysis using timed petri nets. *Journal of Human-Robot Interaction*, 1(1).

[Chetouani, 2014] Chetouani, M. (2014). Role of inter-personal synchrony in extracting social signatures: Some case studies. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, pages 9–12. ACM.

[Clark, 1996] Clark, H. H. (1996). *Using language*, volume 1996. Cambridge university press Cambridge.

[Clark and Krych, 2004] Clark, H. H. and Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81.

[Dong et al., 2007] Dong, W., Lepri, B., Cappelletti, A., Pentland, A. S., Pianesi, F., and Zancanaro, M. (2007). Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 271–278. ACM.

[Gibson, 2003] Gibson, D. R. (2003). Participation shifts: Order and differentiation in group conversation. *Social forces*, 81(4):1335–1380.

[Goffman, 1955] Goffman, E. (1955). On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18(3):213–231.

[Goodwin, 1981] Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. Academic Press New York.

[Grice, 1991] Grice, P. (1991). *Studies in the Way of Words*. Harvard University Press.

[Jonsdottir and Thórisson,] Jonsdottir, G. R. and Thórisson, K. R. A distributed realtime dialogue architecture for dynamically learning polite human turntaking.

[Kilpatrick, 1986] Kilpatrick, P. (1986). Turn and control in puerto rican spanish conversation.

[Knapp, 2012] Knapp, M. L. (2012). *Nonverbal communication in human interaction*. Cengage Learning.

[Kose-Bagci et al., 2008] Kose-Bagci, H., Dautenhahn, K., and Nehaniv, C. L. (2008). Emergent dynamics of turn-taking interaction in drumming games with a humanoid robot. In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, pages 346–353. IEEE.

[LaFrance, 1982] LaFrance, M. (1982). Posture mirroring and rapport. *Interaction rhythms: Periodicity in communicative behavior*, pages 279–298.

[Lévy et al., 2012] Lévy, B., Bloch, G., Assayag, G., et al. (2012). Omaxist dialectics. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 137–140.

[Mehrabian, 1977] Mehrabian, A. (1977). *Nonverbal communication*. Transaction Publishers.

[Novielli et al., 2010] Novielli, N., de Rosis, F., and Mazzotta, I. (2010). User attitude towards an embodied conversational agent: Effects of the interaction mode. *Journal of Pragmatics*, 42(9):2385–2397.

[Pecune et al., 2014] 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 Intelligent Virtual Agents at IVA*, pages 336–342.

[Raux and Eskenazi, 2009] Raux, A. and Eskenazi, M. (2009). A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 629–637. Association for Computational Linguistics.

[Ravenet et al., 2015] Ravenet, B., Cafaro, A., Biancardi, B., Ochs, M., and Pelachaud, C. (2015). Conversational behavior reflecting interpersonal attitudes in small group interactions. In *Intelligent Virtual Agents*, pages 375–388. Springer.

[Sacks et al., 1974] Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735.

[Simpson and Galbo, 1986] Simpson, R. J. and Galbo, J. J. (1986). Interaction and learning: Theorizing on the art of teaching. *Interchange*, 17(4):37–51.

[Ten Bosch et al., 2005] Ten Bosch, L., Oostdijk, N., and Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1):80–86.

[Thórisson, 2002] Thórisson, K. R. (2002). Natural turn-taking needs no manual: Computational theory and model, from perception to action. *Multimodality in language and speech systems*, 19.

[Varni et al., 2015] Varni, G., Avril, M., Usta, A., and Chetouani, M. (2015). Syncpy: a unified open-source analytic library for synchrony. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, pages 41–47. ACM.