Modèles Probabilistes pour l'Interaction entre agents

Kevin Sanlaville[1,2], Gérard Assayag[1], Frédéric Bevilacqua[1], and Catherine Pelachaud[2]

[1]UMR STMS (Ircam, CNRS, UPMC), 1, place Stravinsky, 75004 Paris

[2]LTCI-CNRS, Telecom-Paristech, Université Paris-Saclay, 75013 Paris

ABSTRACT

Dans un contexte d'interaction Humain-Machine, notre objectif est l'élaboration d'un modèle probabiliste d'interaction générique et vraisemblable capable de commander à la fois un Agent Conversationnel Animé (ACA) dans un cadre d'interaction et un Agent Musical Créatif (AMC) dans un contexte d'improvisation musicale.

1. CONTEXTE SCIENTIFIQUE

Dans le cadre du projet SENSE du Labex SMART [1], notre projet de thèse a pour but de réaliser un modèle d'interaction homme-machine adaptable et générique. Nous avons choisi pour cela de nous intéresser aux mécanismes non-verbaux de régulation de la conversation (en anglais Turn-Taking mechanisms). Nous chercherons ensuite à appliquer ce modèle à deux cas, les Agents Conversationnels Animés (ACA) [14] et les Agents Improvisateurs Musicaux (AIM) [13], afin de prouver la généricité de ce modèle. Un tour dans une conversation peut être défini comme le moment entre la prise de parole et son arrêt, qui peut être consensuel ou forcé [9]. Les mécanismes de régulation de la parole sont émis et perceptibles chez le locuteur comme l'interlocuteur et emploient également les mécanismes non-verbaux [12]. Ces signaux sont essentiels à la conversation et garantissent sa fluidité [19]. Alors que certains modèles psychologiques, à l'image du modèle de Sacks [18] supposent que l'organisation en tour survient naturellement dans la conversation et que les tours surviennent à la succession les uns des autres, Clark [3] considère que la notion de tour est émergente. Il ajoute que cette notion prend racine dans l'interaction entre les comportements de parole et d'écoute [4]. En accord avec cette description, nous avons conçu un modèle de conversation axé autour de ces comportements de parole et d'écoute que nous présenterons dans la prochaine section, avant de détailler nos plans pour une expérimentation.

2. OBJECTIFS

Les buts de notre modèle sont multiples. Cherchant à modéliser les comportements non-verbaux qui régulent la conversation (qui peuvent signifier l'écoute, un changement de locuteur, le désir pour le locuteur présent de conserver la parole, etc.), nous souhaitons que notre modèle soit adaptable. Ainsi, il sera capable de modéliser non seulement des comportements différents pour un contexte donné, mais il sera également à même de modéliser plusieurs contextes. Notre projet de thèse s'intéressant à la fois à la modélisation de la conversation et de l'improvisation musicale libre, nous souhaitons que ce modèle soit générique. En lui faisant apprendre les comportements à partir de données réelles, nous souhaitons faire en sorte que notre modèle décrive correctement ces deux contextes et leur spécificités (alors que dans certaines conversation le but est de parler chacun son tour, le but de la plupart des musiciens est de jouer en même temps, dans la musique certains gestes sont purement performatifs et doivent être ignorés, etc.). Notre troisième but est enfin la vraisemblance. Nous souhaitons que les comportements générés soient considérés comme similaires à la réalité par des observateurs extérieurs et nous comptons parvenir à cet objectif en faisant émerger des comportements de chaque agent une synchronie comme définie comme adaptation dynamique et réciproque de la structure temporelle des comportements entre partenaires interactifs [5].

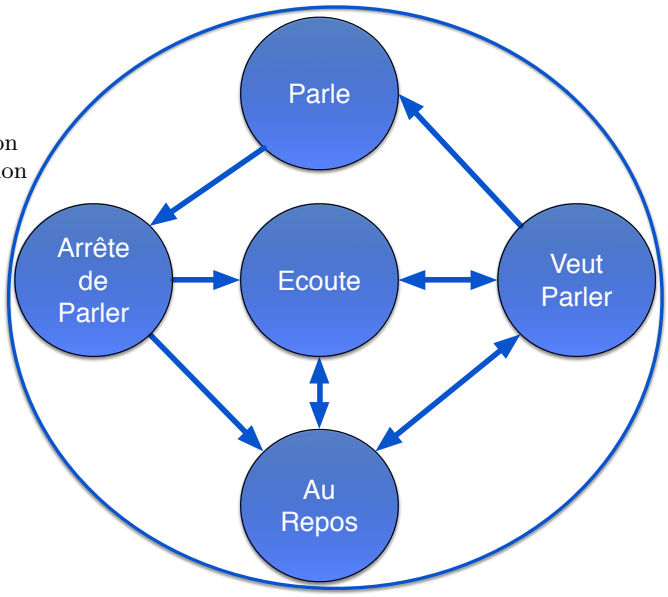

Figure 1: Modèle de conversation du point de vue des comportements d'un agent

3. POSITIONNEMENT

3.1 Structure du modèle

Le Turn-Taking est un domaine relativement vaste et de nombreux modèles ont déjà été présentés (pour plus d'informations,

[1]http://www.smart-labex.fr/index.php?perma=SeNSE

se référer à [2]). Nous avons choisi d'utiliser un Automate à Etats Finis (AEF) pour modéliser la conversation. L'utilisation d'AEF est présente dans la littérature, car d'après Gibson les comportements conversationnels découlent des états mentaux des locuteurs [7]. Il existe deux types d'approche : soit les AEF modélisent le système entier (comme dans Raux et al. [15] où l'agent a une vue du système et si lui ou un autre agent parle) ou l'AEF modélise les états internes de l'agent (comme dans Jonsdottir et al. [11] où l'AEF décrit si l'agent veut parler, est en train de parler ou écoute). Modéliser la conversation dans son entier s'approcherait des théories de Sacks car les règles conversationnelles y seraient déterminées à l'avance ; nous avons choisi de modéliser les comportement de l'agent en pensant que dans ce comportement les règles conversationnelles émergeraient d'elles-mêmes, en nous appuyant sur l'hypothèse de Goffman qui suppose une relation entre la structure du soi et la structure de l'interaction [8]. Ł'ensemble du modèle peut être observé dans la Figure 1. En nous appuyant sur les thèses de Clark [3], nous supposons que le modèle doit comprendre les actions de parole et d'écoute. Cependant, d'après Sacks, le nombre de participants peut varier au cours de la conversation [18] ; d'où notre ajout d'un état "Au Repos" pour décrire la situation initiale ou personne ne participe encore à la conversation et si un agent se désintéresse de la conversation et n'envoie donc plus de signaux d'écoute. Cet état est à la fois l'état initial et final de notre AEF. Enfin, les phases de prise et de libération de la parole font aussi partie intégrante de la conversation car elles assurent une bonne coordination entre interlocuteurs[19]. Nous avons donc ajouté les états "Veut le Tour" et "Donne le Tour" à notre modèle. Après avoir présenté la structure de notre modèle, nous allons à présent en détailler le mode de transition.

3.2 Modèle probabiliste

Nous avons déjà mentionné que nous utilisons des modèles à états. Cependant, il nous fallait trouver comment les agents allaient passer d'un comportement à un autre, c'est à dire d'un état à un autre. Grice affirmant que nous inférons les comportements conversationnels à partir des croyances que nous avons à propos d'elles [10], nous avons choisi d'implémenter notre modèle à états à l'aide de Modèles de Markov Cachés (Hidden Markov Model, HMM) [1], où les états du modèles seraient les états du HMM et les probabilité de transition seraient apprises à partir de données réelles. Cependant, les décisions de prise de parole ne sont pas seulement issues du comportement présent de l'agent mais aussi de celles des autres agents. Ce pourquoi nous avons décidé de modéliser l'ensemble des agents par un modèle d'influence [6]. 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. 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.

4. VALIDATION

Afin de vérifier la pertinence de l'usage des modèles d'influence,

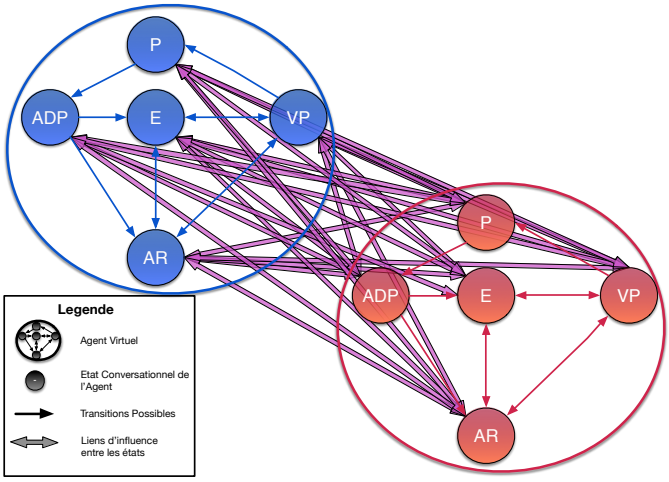

Figure 2: Représentation du modèle d'influence

nous avons utilisé quatre agents GRETA dans un environnement créé dans UNITY [2] (cf. travaux de Ravenet et al. [16]), nous avons généré des traces que nous avons utilisé pour apprendre à notre modèle d'influence. Les observations que nous avons tirées de cette première expérimentation sont de plusieurs natures :

- Nous avons observé que les matrices de transition de tous les agents convergeaient vers des valeurs stables
- Nous avons également observé que les temps de parole étaient du même ordre que dans le cas de modèles originaux
- Nous avons observé que le modèle répliquait les comportements du modèle quels qu'ils soient, i.e. que si les traces à partir duquel le modèle apprenait montraient une répartition égale du temps de parole, les traces générées respectaient cette organisation, et de même si un agent monopolisait la parole et que les autres le laissaient faire, les traces générées suivaient cette répartition du temps de parole

La prochaine étape de notre projet de thèse sera de lancer une expérimentation où nous demanderons à des sujets naïfs d'évaluer la vraisemblance des interactions que nous aurons enregistrées en les comparant au modèle de Ravenet et al. et à des HMMs simples. Nous utiliserons également les outils développés par Varni et al. [20] 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.

5. PERSPECTIVES

Les perspectives applicatives sont multiples. Une première application évidente sera l'application de notre système au système OMaX [13], qui aujourd'hui encore a besoin d'un opérateur humain pour signifier au système quand il doit jouer ou non. Une autre application possible pourrait être une application thérapeutique. En effet, il a été prouvé que des patients souffrant de phobie réagissaient mieux et conservaient plus longtemps les effets de leur thérapie lorsqu'ils

[2]http://unity3d.com

étaient confrontés à leur phobie dans un contexte de réalité virtuelle [17]. Il semble envisageable d'utiliser nos agents afin de permettre à des personnes souffrant de difficultés à interagir avec d'autres êtres humains (troubles du spectre autistique, schizophrénie) de s'entraîner à interagir sans le risque d'être effrayé par le contact d'un autre être humain et en maîtrisant mieux les paramètres de l'interaction.

6. ACKNOWLEDGMENTS

Ce travail été effectué dans le cadre du Labex SMART (ANR-11-LABX-65) et a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir portant la référence ANR-11-IDEX-0004-02 et le projet DYCI2 portant la référence ANR-14-CE24-0002-01.

7. REFERENCES

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

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

[3] Herbert H Clark. *Using language*, volume 1996. Cambridge university press Cambridge, 1996.

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

[5] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *Affective Computing, IEEE Transactions on*, 3(3):349–365, 2012.

[6] Wen Dong, Bruno Lepri, Alessandro Cappelletti, Alex Sandy Pentland, Fabio Pianesi, and Massimo Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 271–278. ACM, 2007.

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

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

[9] Charles Goodwin. *Conversational organization: Interaction between speakers and hearers*. Academic Press New York, 1981.

[10] Paul Grice. *Studies in the Way of Words*. Harvard University Press, 1991.

[11] Gudny Ragna Jonsdottir and Kristinn R Thórisson. A distributed realtime dialogue architecture for dynamically learning polite human turntaking.

[12] Mark L Knapp. *Nonverbal communication in human interaction*. Cengage Learning, 2012.

[13] Benjamin Lévy, Georges Bloch, Gérard Assayag, et al. Omaxist dialectics. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 137–140, 2012.

[14] Radosław Niewiadomski, Mohammad Obaid, Elisabetta Bevacqua, Julian Looser, Le Quoc Anh, and Catherine Pelachaud. Cross-media agent platform. In *Proceedings of the 16th International Conference on 3D Web Technology*, pages 11–19. ACM, 2011.

[15] Antoine Raux and Maxine Eskenazi. 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, 2009.

[16] Brian Ravenet, Angelo Cafaro, Beatrice Biancardi, Magalie Ochs, and Catherine Pelachaud. Conversational behavior reflecting interpersonal attitudes in small group interactions. In *Intelligent Virtual Agents*, pages 375–388. Springer, 2015.

[17] Barbara Olasov Rothbaum, Larry Hodges, Page L Anderson, Larry Price, and Samantha Smith. Twelve-month follow-up of virtual reality and standard exposure therapies for the fear of flying. *Journal of consulting and clinical psychology*, 70(2):428, 2002.

[18] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735, 1974.

[19] Louis Ten Bosch, Nelleke Oostdijk, and Lou Boves. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1):80–86, 2005.

[20] Giovanna Varni, Marie Avril, Adem Usta, and Mohamed Chetouani. 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, 2015.