From Analysis to Modeling of Engagement as Sequences of Multimodal Behaviors

Soumia Dermouche, Catherine Pelachaud
CNRS UMR 7222, ISIR, Sorbonne Universités
4 Place Jussieu Paris, France
{dermouche, pelachaud}@isir.upmc.fr

Abstract

In this paper, we present an approach to endow an Embodied Conversational Agent with engagement capabilities. We relied on a corpus of expert-novice interactions. Two types of manual annotation were conducted: non-verbal signals such as gestures, head movements and smiles; engagement level of both expert and novice during the interaction. Then, we used a temporal sequence mining algorithm to extract non-verbal sequences eliciting variation of engagement perception. Our aim is to apply these findings in human-agent interaction to analyze user's engagement level and to control agent's behavior. The novelty of this study is to consider explicitly engagement as sequence of multimodal behaviors.

Keywords: Engagement, Non-verbal behavior, ECA, Temporal Sequence Mining, Human-agent interaction

1. Introduction

Embodied Conversational Agents (ECA) are virtual characters that can interact with a user. Today, ECAs are increasingly being integrated in our everyday life, for example, for training, social coaching, and science teaching (Graesser et al., 2007). Our work is part of the H2020 European project ARIA-VALUSPA (Valstar et al., 2016) that aims to build an ECA able to play the role of an expert and to share its domain knowledge with a novice user. In this project, we focus on an important aspect of human-agent interaction, namely, engagement that ensures the interaction to go on. A survey of engagement definition in human-agent interaction is given in (Glas and Pelachaud, 2015). Engagement can be defined as: *"the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction"*(Poggi, 2007). Engagement is also defined as "the process by which participants involved in an interaction start, maintain and terminate an interaction" (Sidner et al., 2005). Engagement is not measured from single cues, but rather from several cues that arise over a certain time window (Peters et al., 2005).

The goal of this work is twofold: (1) user's engagement detection: the ECA should be able to detect, in real time, the engagement level of the user. (2) ECA's engagement modeling: the ECA should adapt its behavior in order to maintain the desired level of engagement during the interaction. Specifically, this work investigates what are the multimodal behaviors that participate to a change of perception of the engagement level in a human-agent interaction. For this, we rely on sequence mining algorithms to associate user's and agent's non-verbal behaviors with different engagement levels.

Several works focused on associating verbal and non-verbal behaviors with engagement in human-agent interaction, but most of them are limited to few signals such as smiling and head nod (Allwood and Cerrato, 2003; Castellano et al., 2009). In our work, we consider a set of non-verbal modalities (gesture, head movement and directions, smiling, etc.) jointly with their temporal synchronization (order, starting time and duration).

Our study is performed on the NoXi dataset, a corpus of expert-novice interaction (Cafaro et al., 2017), that we have manually annotated according to the engagement levels of both expert and novice. The use of sequence mining allowed us discovering relevant patterns for different engagement levels.

2. Related Works

During the last decades, engagement modeling has gained increasing attention due to the growing number of conversational agents and the important role that engagement plays in human-agent interaction. Engagement can be expressed by both verbal and non-verbal behaviors. Engagement can be directly linked, for example, to prosodic features (Yu et al., 2004) and verbal alignment behaviors (Pickering and Garrod, 2004).

Other studies have reported that smiling (Castellano et al., 2009) and head nod can provide information about the participant's engagement (Allwood and Cerrato, 2003). Gaze is also an important cue of engagement level (Sidner et al., 2003; Peters et al., 2005; Nakano, Yukiko I. and Ishii, 2010), for example, looking at the speaking partner can be interpreted as a cue of engagement, while looking around the room may indicate the intention to disengage. Moreover, a correlation has been found between engagement and several body postures (Mota and Picard, 2003; Sanghvi et al., 2011). In short, engagement can be conveyed by multimodal behaviors (Sidner et al., 2003).

Results from (Ivaldi et al., 2017) confirmed the relationship between attitudes and engagement in human-robot interaction. For instance, more user is extrovert, he tend to more talk to the robot. Also a negative attitude towards robots have been correlated with less gaze at the robot's face. Culture is an important aspect to take in account when modeling engagement for virtual agents (Yu et al., 2016; Matsumoto, 2006). Yu *et al.* found that in American culture, more smiles represents more engagement, while in Chinese culture, similes are less related to engagement. Another

Self_adaptor Iconic Beat Deictic Metaphoric

Figure 1: Examples of rest gestures.

example is that Arabs gaze (gaze and speech are the main social signals used to evaluate engagement (Sidner et al., 2010)) much longer and more directly at their partners than do Americans (Matsumoto, 2006).

3. Multimodal Corpus Representation

In this section, we present the NoXi corpus, as well as the tool we have used for annotation (NOVA). We also present our annotation scheme for the non-verbal behavior and engagement.

3.1. NoXi

This work is part of the H2020 project ARIA-VALUSPA (Valstar et al., 2016) (*Artificial Retrieval of Information Assistants - Virtual Agents with Linguistic Understanding, Social skills and Personalized Aspects*). In this project, a database of multilingual natural dyadic interactions, named NoXi (Cafaro et al., 2017), has been collected. NoXi is publicly available through a web interface[1]. NoXi provides spontaneous interactions that involve an expert and a novice discussing about a given topic (e.g.; sports, politics, videogames, travels, music, etc.). The dataset contains over 25 hours of dyadic interactions spoken in multiple languages (mainly English, French, and German). In this work, we use the French part of NoXi database which is composed of 21 sessions. The total duration of all these sessions is 7 hours and 25 minutes.

3.2. NOVA

In the context of the ARIA-VALUSPA project, a graphical tool named NOVA (Baur et al., 2015) has been developed to review and annotate the recorded data [2]. NOVA allows exploring richer data such as skeleton or face streams and by proposing various annotation schema (discrete or continuous). We use NOVA as annotation tool.

3.3. Manual Annotations of NoXi Corpus

Table 1 summarizes the multimodal behaviors that we manually annotated by adapting the MUMIN multimodal coding scheme (Allwood et al., 2007). We use a discrete annotation scheme to label body behavior (e.g., gesture, gaze direction and head movement) and continuous scale for engagement annotation. The manual annotations that we have realized on NoXi corpus will be publicly available through a Web interface[3].

- **Conversation states**
  We annotate four conversation states: both interlocutors speak (BOTH), expert speaks (EXPERT), novice speaks (NOVICE) or no one speaks (NONE).

- **Facial display**
  For facial behavior, we considered: gaze, head movement/direction, smile, and eyebrow movement.

- **Gesture**
  Based on the taxonomies proposed by McNeill (1992), we annotate five categories of gestures: iconics, metaphorics, deictics, beats, and adaptors defined as follows:

  1. Iconics: describe concretely the object that the discourse is presenting.
  2. Metaphorics: in contrary to iconic gestures, these gestures illustrate the speech in an abstract way.
  3. Deictics: point to a location in space, for example, an object a place or a concrete direction),
  4. Beats: do not include semantic information, they are characterized by their simplicity and receptivity.
  5. Adaptors: serve to satisfy bodily needs like scratching.

  Examples of different types of gestures are showed in Figure 1. We also include hand rest positions that can indicate communicator's status and attitude (Allwood et al., 2007). We consider several positions (see Figure 2): arms crossed, hands together, hands in pockets, hands behind back, akimbo (hands on hips), along body (arms are stretched down along the body) etc.

- **Engagement**
  Based on Poggi's definition of engagement, we have continuously annotated the engagement of both expert

[1] https://noxi.aria-agent.eu/
[2] https://github.com/hcmlab/nova
[3] https://noxi.aria-agent.eu/

(1) (2) (3) (4) (5) (6)

Figure 2: Examples of rest poses: (1) hands_crosseddown, (2) hand_inpocket, (3) hand_along, (4) hands_crossed, (5) hands_onhips, (6) hands_behaind.

Table 1: Annotation scheme for the non-verbal behaviors and engagement annotations in NoXi.

Tier (modality)	**Labels**
Conversation states	NONE — EXPERT — NOVICE— BOTH
Head movements	NOD — SHAKE
Head direction	FORWARD — BACK — UPWARDS — DOWNWARDS — SIDEWAYS — SIDE_TILT
Smiles	SMILE
Eyebrow movements	RAISED — FROWN
Gaze direction	TOWARDS_INTERLOCUTOR — UP — DOWN — SIDEWAYS — OTHER
Gestures	ICONIC — METAPHORIC — DEICTIC — BEAT — ADAPTOR
Hand rest positions	ARMS_CROSSED — HANDS_TOGETHER — HANDS_IN_POCKETS — HANDS_BEHIND_BACK — AKIMBO — ALONG_BODY
Engagement	STRONGLY DISENGAGED ... STRONGLY ENGAGED

and novice. To reduce complexity and facilitate the task of continuous annotation, we have defined five levels to annotate changes in the perception of engagement: strongly disengaged, partially disengaged, neutral, partially engaged, strongly engaged. In order to avoid content biases from the verbal behavior when annotating engagement, we have filtered it out, for both expert and novice by applying a Pass Hann Band Filter. In this way, the speech kept the prosodic information without intelligibility of its verbal content.

4. Corpus Analysis

In this section, we present an analysis of the manual annotation. Each single modality (gesture, rest positions, engagement, etc.) has been annotated by one annotator. The inter-annotator agreement (Cohen's Kappa) is greater than 0.5 for all modalities which means that there is a high level of agreement between annotators. Because of space limitation, we only present results about gesture and rest hand positions.

- **Gesture**
  Table 2 shows the number of gestures produced by the expert and the novice. As it can be seen, the expert produces 4 times more gestures (1223) than the novice (293). During the interaction, the expert controls the discussion topic: he holds the floor and he produces more gestures to explain and illustrate his topic. Gestures were mainly either iconics or metaphorics. These gesture types contribute to the perception of higher level of competence according to (Maricchiolo et al., 2009).

- **Rest arms and hand positions**
  The number of arms positions produced by the expert is much more important than that of the novice (cf. Table 3). This can be explained as the novice is mainly a listener and keeps his rest position much longer (mean duration 32.2 seconds) than the expert (mean duration 10.7 seconds).

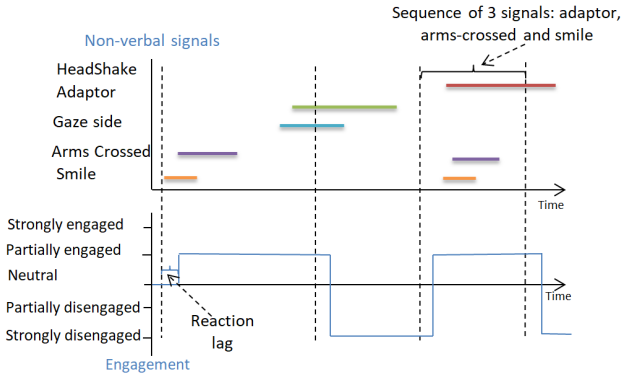

Figure 3: Non-verbal behavior segmentation based on engagement variation. The result is a set of non-verbal signal sequences for each engagement level. For each level, we apply HCApriori to extract the most relevant patterns representing this engagement level.

Table 2: Number of gestures produced by expert and novice.

	Iconics	Metaphorics	Deictics	Beats	Adaptors	Total
Expert	211	158	67	704	83	1223
Novice	49	19	11	105	109	293

Table 3: Number of rest positions for expert and novice.

	Crossed	Together	In_pockets	Behind_back	Akimbo	Along body	Total
Expert	212	486	189	93	48	289	1317
Novice	82	177	214	32	53	54	612

5. Sequence-based Engagement Modeling

Human behaviors are naturally multimodal and sequential: we interact with each other through multiple communication channels (speech, gaze, gesture, etc.). Moreover, these behaviors are temporally coordinated: what behavior we will display next depends, among other phenomena, on our behavior at the present moment and on the other's behavior.

The goal of the present study is to understand how those behaviors are coordinated at critical moments, the sequential patterns they exhibit and their association with different engagement levels. To capture both sequentiality and temporality, we rely on temporal sequence mining, a data mining technique that considers the temporal information like starting time and the duration of signals, a key element in behavior modeling. The ECA should display behaviors at the right moment with the right duration in order to convey a given level of engagement.

A range of temporal sequence mining algorithms exist like HCApriori (Dermouche and Pelachaud, 2016), QTIPrefixSpan (Guyet and Quiniou, 2011) and PESMiner (Ruan et al., 2014). In this work, we rely on HCApriori because it demonstrated a superiority over the state-of-the-art in terms of pattern extraction accuracy and running time (Dermouche and Pelachaud, 2016). In order to prepare a sequence database for HCApriori, we have segmented the non-verbal behaviors based on engagement variations (cf. Figure 3). We took into account the reaction lag of annotators in the continuous annotations by shifting back 2 seconds each of the annotations, as recommended in (Mariooryad and Busso, 2013). For each engagement level, we consider the sequence of non-verbal signals that simultaneously appeared with this level (cf. Figure 3). Thus, we build five datasets of non-verbal signal sequences representing the five engagement levels. Table 4 summarizes the number of sequences we obtain for each engagement level for expert and novice. Finally, we have applied HCApriori to extract temporal patterns (frequent sub-sequences) of nonverbal signals expressing the five engagement levels.

Figure 4 shows a pattern extracted with HCApriori algorithm representing a strong engagement level. This pattern can be interpreted as follows: 0.9 second before the annotator perceives a strong engagement level, the expert smiles to the novice for 1.4 seconds while nodding his head. Then he produces an akimbo gesture (hands on hips) meanwhile he nods and continues smiling. Smiling and head nod have already been reported as being engagement indicators (Castellano et al., 2009; Allwood and Cerrato, 2003). Table 5 gives two examples of extracted patterns (P_1, P_2) for each engagement levels of expert. The starting and the duration of each signal are respectively given between two parentheses. While head nod has been reported as a cue of engagement in several works, our approach associates it with a partial disagreement if it is preceded by a frown and associates it with a strong engagement if it coincides only with a smile. This indicates that the perception of non-verbal signal may change with respect to the signals occurring before and after it.

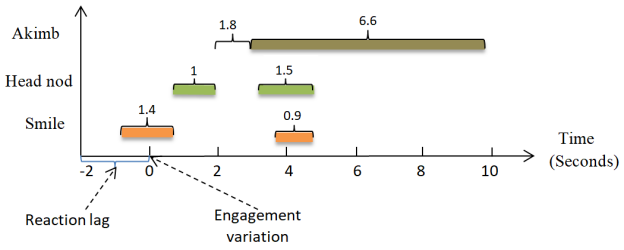

Figure 4: Pattern example representing a strong engagement level.

In Table 6, we compute the percentage of some non-verbal signals that occurred in the patterns of expert for the five engagement levels. Here are some conclusions that we can draw from this table:

- Smile signal occurred in about 66% of the patterns representing a strongly engaged level. Within this level of engagement, the mean starting time of smile is -0.05 seconds, this means that 0.05 seconds before the perception a strong engagement, the expert smiles. In other words, the perception of strong engagement is triggered by the expert's smile. On the other hand, for the partially disengaged level, the mean starting time of smile is 1.5 seconds: smile is produced 1.5 seconds after the perception of a partial engagement.

- Adaptor gestures and frown were also more frequently present in patterns characterizing a strongly disengaged level with a percentage of 25% and 33% respectively.

- Although head nod is usually reported as an indicator of engagement, we found that this signal is more representative of disengagement (33% occurrence in strong disengagement level). This suggests that the

Table 4: Number of sequences of each engagement level for both expert and novice.

	Strongly disengaged	Partially disengaged	Neutral	Partially engaged	Strongly engaged	Total
Expert	48	373	373	561	126	1481
Novice	116	432	509	558	64	1679

Table 5: Some examples of extracted patterns for the five engagement levels.

Engaged level	Pattern example
Strongly disengaged	P_1= Eyebrow down (2, 1.48), P_2= Adaptor(4.08, 2.12)
Partially disengaged	P_1=Eyebrow down(-1.6, 5) Head nod(5.1, 2.7), P_2=Arms crossed(-1.68, 4.12) EyebrowUp (-1.36, 2.64)
Neutral	P_1=Along_body(-1.04, 9.5) Beat(0.37, 3.5), P_2= Beat(-1.4, 2.28) Smile (1.56, 1) Nod(2.52,1.72)
Partially engaged	P_1=Metaphoric (-1.8, 1.75) Iconic(-0.3,2.5), P_2=Iconic(-1.25,2.22) Iconic(2,5)
Strongly engaged	P_1=Smile(-0.76, 3.64) Head nod (0.24, 0.36), cross (-1.14, 13.77), P_2= Smile(-1.08,1.36) Smile (1.64,1.4)

Table 6: Percentage of some non-verbal signals that occurred in the patterns of expert for the five engagement levels.

	Strongly disengaged	Partially disengaged	Neutral	Partially engaged	Strongly engaged
Smile	0%	4%	16%	20%	66%
Eyebrowdown	33%	20%	25%	18%	8%
Nod	33%	20%	22%	24%	25%
Adaptor	25%	10%	11%	3%	0%

perception of non-verbal signal change according to its context. For example, nodding while smiling and performing an adaptor gesture was associated with a partial disengaged level.

6. Conclusion

In this paper, we presented a sequence-mining based approach toward engagement modeling from a corpus of expert-novice interactions. Sequence mining allowed us to extract relevant patterns associated to five engagement levels. While a part of our results perfectly supports previous works, some of our findings are complementary to the current state-of-the-art. This demonstrates that temporal characteristics, like starting time and duration of behaviors, are essential to studying engagement.

Our future purpose is to apply sequence mining results in human-agent interaction: (1) using expert patterns to model the desired engagement level of an ECA during the interaction. (2) Exploiting the patterns representing novice engagement for interpreting user's non-verbal behaviors in real-time and associate it with different engagement variations for allowing the agent to react accordingly.

7. Acknowledgements

Funded by European Union Horizon 2020 research and innovation programme, grant agreement No 645378.

8. Bibliographical References

Allwood, J. and Cerrato, L. (2003). A study of gestural feedback expressions. In *First Nordic Symposium on Multimodal Communication*, pages 7–22.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The MUMIN annotation scheme for feedback, turn management and sequencing. *International Journal of Language Resources and Evaluation*, pages 1–18.

Baur, T., Mehlmann, G., Damian, I., Lingenfelser, F., Wagner, J., Lugrin, B., André, E., and Gebhard, P. (2015). Context-Aware Automated Analysis and Annotation of Social Human–Agent Interactions. *ACM Transactions on Interactive Intelligent Systems*, 5(2):1–33.

Cafaro, A., Wagner, J., Baur, T., Dermouche, S., Torres, M. T., Pelachaud, C., Andr, E., and Valstar, M. (2017). The NoXi Database : Multimodal Recordings of Mediated Novice-Expert Interactions. In *ICMI'17,*, pages 350–359, Glasgow, Scotland. ACM.

Castellano, G., Pereira, A., Leite, I., Paiva, A., and McOwan, P. W. (2009). Detecting user engagement with a robot companion using task and social interaction-based features. *Proceedings of the 2009 international conference on Multimodal interfaces - ICMI-MLMI '09*, (January 2009):119.

Dermouche, S. and Pelachaud, C. (2016). Sequence-based multimodal behavior modeling for social agents. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016*, pages 29–36, Tokyo, Japan. ACM.

Glas, N. and Pelachaud, C. (2015). Definitions of engagement in human-agent interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, pages 944–949.

Graesser, A., Chipman, P., King, B., Mcdaniel, B., and Mello, S. D. (2007). Emotions and Learning with AutoTutor. *13th International Conference on Artificial Intelligence in Education (AIED 2007)*, pages 569–571.

Guyet, T. and Quiniou, R. (2011). Extracting temporal

patterns from interval-based sequences. In *International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 1306–1311.

Ivaldi, S., Lefort, S., Peters, J., Chetouani, M., Provasi, J., and Zibetti, E. (2017). Towards Engagement Models that Consider Individual Factors in HRI : On the Relation of Extroversion and Negative Attitude Towards Robots to Gaze and Speech During a Human â Robot Assembly Task Experiments with the iCub humanoid. *International Journal of Social Robotics*, 9(1):63–86.

Maricchiolo, F., Gnisci, A., Bonaiuto, M., and Ficca, G. (2009). Effects of different types of hand gestures in persuasive speech on receivers' evaluations. *Language and Cognitive Processes*, 24(2):239–266.

Mariooryad, S. and Busso, C. (2013). Analysis and Compensation of the Reaction Lag of Evaluators in Continuous Emotional Annotations. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013)*, pages 85–90.

Matsumoto, D. (2006). Culture and nonverbal behavior. *Handbook of nonverbal communication*, pages 219–236.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Uni- versity of Chicago press.

Mota, S. and Picard, R. W. (2003). Automated Posture Analysis for Detecting Learner's Interest Level. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 5:1–6.

Nakano, Yukiko I. and Ishii, R. (2010). Estimating User's Engagement from Eye-gaze Behaviors in Human-agent Conversations. In *The 5th International Conference on Intelligent User Interfaces*, pages 139—-148. ACM.

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., and Poggi, I. (2005). Engagement Capabilities for ECAs. *AAMAS'05 workshop Creating Bonds with ECAs*.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and brain sciences*, 27(2):169–190; discussion 190–226.

Poggi, I. (2007). *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*.

Ruan, G., Zhang, H., and Plale, B. (2014). Parallel and Quantitative Sequential Pattern Mining for Large-scale Interval-based Temporal Data. In *2014 IEEE International Conference on Big Data (BigData 2014)*, pages 32–39.

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., and Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. *Proceedings of the 6th international conference on Human-robot interaction - HRI '11*, page 305.

Sidner, C. L., Lee, C., and Lesh, N. (2003). Engagement by Looking: Behaviours for Robots when Collaborating with People. *Proceedings of DiaBruck (the 7th Workshop on Semantics and Pragmatics of Dialogue)*, pages 123–130.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164.

Sidner, C. R., Ponsleur, B., Holroyd, A., and L., C. (2010). Recognizing engagement in human-robot interaction. *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382.

Valstar, M., Baur, T., and Wagner, J. (2016). Ask Alice: An Artificial Retrieval of Information Agent. In *ICMI'17*, pages 419–420.

Yu, C., Aoki, P. M., and Woodruff, A. (2004). Detecting User Engagement in Everyday Conversations. *Science*, page 4.

Yu, Z., He, X., Black, A. W., and Rudnicky, A. I. (2016). User Engagement Study with Virtual Agents Under Different Cultural Contexts. In *Intelligent Virtual Agents - 16th International Conference, IVA2016*, pages 364–368, Los Angeles, CA, USA.