Specifying Cooperation between Modalities in Lifelike Animated Agents

ABRILIAN Sarkis[1], BUISINE Stéphanie[1], RENDU Christophe[1], MARTIN Jean-Claude[1&2]
martin@limsi.fr http://www.limsi.fr/Individu/martin/research/projects/lea/

(1) LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
(2) LINC-Univ. Paris 8, IUT de Montreuil, 140 Rue de la Nouvelle France, 93100 Montreuil, France

Abstract

In this paper we introduce the Limsi Embodied Agent project which tackles both the need to ground agent's behavior on video-taped annotations of application dependent human behavior and the granularity of the language for specifying the agent multimodal behavior.

Keywords

Verbal and non-verbal behavior, tools for scripting behavior, pedagogical agents.

Introduction

There is still a lack of appropriate and global answers to the question of the "lifelikeness" and of "believability" of animated agents. The specification of multimodal behavior of animated agents is often based on knowledge extracted from the literature in several domains such as Psychology, Sociology and Linguistics. As partly suggested by (Kipp 2001a ; Cassell et al. 2001b), we believe that in order to be really lifelike, multimodal behavior of agents needs to be grounded in experimental studies in the same application context (i.e. the multimodal behavior of pedagogical agents should be based on video recording and annotation of teacher's behavior in "similar" settings). In this paper, we describe how we intend to use such an experimental approach with the Limsi Embodied Agent (LEA).

But how do we go from annotating human multimodal behavior to specifying the behavior of an agent? Existing specification languages are mostly dedicated either to low-level monomodal specification (i.e. angry facial expression) or to amodal "higher" level specifications which are translated into monomodal features (i.e. angry behavior generating facial expression, intonation, gaze…).

In this paper we describe how we will define, within the LEA project, an intermediate level of specification based on types of cooperation between communicative modalities which can be useful for fine-grain specification of multimodal communicative behavior based on video corpus annotation (Martin et al. 2001).

Grounding multimodal behavior on video annotation

Annotating human multimodal behavior

Following previous work on manual annotation of video-taped human multimodal behavior, we have developed tools making easier the annotation and the computation of behavioral metrics. We have defined a grammar for such annotations (a XML DTD). According to this grammar, these annotations are composed of several sections. A first section describes the features the subject is referring to in the corpus (ie. objects drawn on the blackboard in the case of a teacher). Each of the following sections contains the annotation of a multimodal segment, itself composed of several sub-sections potentially including annotation of references to objects in each modality such as speech, hand gesture, gaze.

Computing metrics of human multimodal behavior

A Java software has been developed in order to parse these annotations of human multimodal behavior and to compute behavioral metrics (Martin et al. 2001). It follows the following steps:

- Parse the file containing the annotation and build internal representation
- Assign a « salience » value to each (object, reference) couple according to rules such as « *if the referent contains the fully specified name of the object, assign value 1.0 to the salience value* »
- Assign a priori fixed values to weights for each modality

- Compute the average salience value in each single multimodal segment across all modalities
- Compute the average salience value for each object across all modalities
- Compute behavioral metrics (complementarity / redundancy rate, equivalence rate)

From human behavior annotation to agent behavior specification

Both the DTD and the software have been already applied to 40 samples taken from several corpora. We are currently studying how to integrate this approach with the Anvil tool (Kipp 2001b) as described in Figure 1. We intend to evaluate such tools on larger corpora and to integrate them in a larger methodology for the analysis of multimodal behavior that we will apply to several domains such as e-learning (Martin et al. 2002).

One long term goal we have is to find an efficient way for establishing a systematic mapping between annotations of human behavior and specifications of the multimodal behavior of the corresponding agent. The resulting behavioral metrics (redundancy/complementarity rate, equivalence rate…) will form the basis of the language we propose for specifying the multimodal behavior of agent that we describe in the next section.

Specifying cooperation between modalities in agent behavior

Granularity level of existing agent specification languages

Existing animated agent specification languages can be compared on the basis of several criteria including the available modalities and the granularity of the specification tags (Table 1). The VHML language (Gustavsson et al. 2001) is used to facilitate the interactions between a virtual agent and the user by featuring one specification sub-language for each modality (GML for gestures, SML for speech, BAML for body, FAML for facial expression) but also specification sub-languages for "higher" amodal levels (EML for emotion, DMML for Dialogue Manager Markup Language). The BEAT project (Cassell et al. 2001c) enables the animation of an avatar by using typed text. It makes use of behavioral "suggestive functions", for example the "Surprising Feature Iconic Gesture Generator" function (movements generated when the avatar encounters surprising information). Behavior selection is achieved by two filters: one for the resolution of conflicts and the other for the priority threshold.

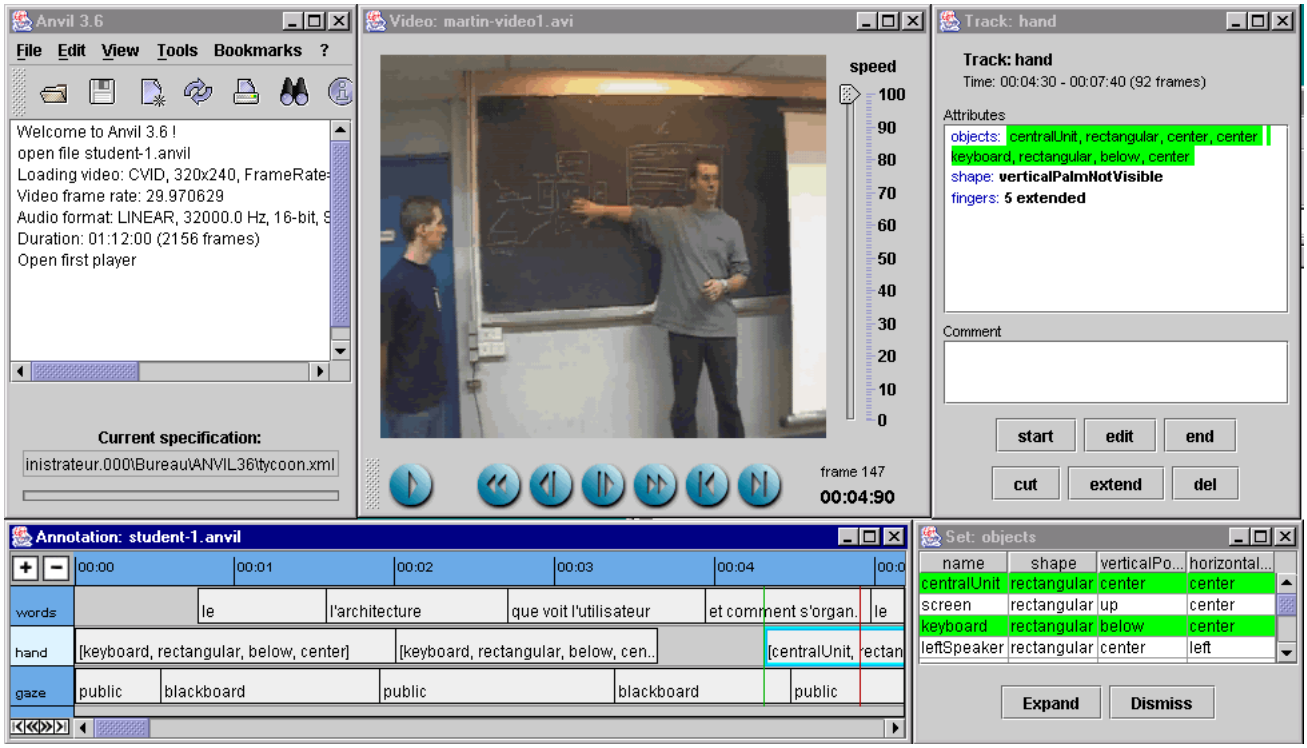

Figure 1: Screenshot of the annotated example. The annotation scheme (lower left window) contains 3 tracks (spoken words, hand gestures, gaze). A new Anvil feature enables the annotation of objects referred in gesture and speech (lower right window) as described in (Martin & Kipp 2002).

	Combi-nation of modalities	Conver-sational Behavior	Enviro-nment	Gaze	Arms & hands gestures	Head gestures	Face	Body posture	Speech		
									Recog-nizer	Synth-esizer	Intona-tion
REA	+	++	-	+	+	+	+	+	+	+	+
BEAT	+	+	-	+	+	-	+	-	+	-	+
VHML	++	+++	-	+	-	+	+	-	-	+	+
Traum	+	++	+++	+	+	+	+	+	+	+	+
Guerrin	++	+	+	-	+	-	+	+	-	+	+
Storyteller	+	+	++	-	+	+	+	+	-	+	++

Table 1: Some animated agent specification languages (-: not used ; +++,++,+: modality more or less used)

BEAT is used in MACK (Cassell et al. 2002) which automatically annotates a text with the following modalities: hand gesture, gaze, eyebrow, body movement and intonation. In the specification language of REA (Cassell et al. 1999) different high-level functions combine several modalities. For example the "Give turn" function trains the hands' relaxation, a glance towards the user and the lifting of the eyebrows. The "Open interaction" function trains the eye to look towards the user, a smile and a head toss.

In (Traum & Rickel 2001) the language is specified to manage interactions between a group of immerged agents in a virtual world, and of which the specifications are in the form of conversational tags: make-contact, break-contact, give-attention, release-attention, start-topic, end-topic. Contrary to the works of (Guerrin et al. 2001) whose rules are related to a specific communication plan between two agents (seller and client), containing tags of a low level (defining specific expressions) and at the same time tags of a high level, which corresponds to combinations of low-level tags, or to works done by (Silva et al. 2001) where a unique active agent, the storyteller, is interacting with a passive user, and of which the specification language is based on four variables: behavior, environment, emotion and time of the day.

Low level specification of monomodal behavior

The current version of our LEA agent is written in Java and parses an XML file containing a sequence of configurations (Table 2). The corresponding screendumps are given in Figure 2.

The current version is thus limited to the manual specification of each single monomodal configuration. The program uses single frame animation (gaze, facial expression, arms, head, body) and speech synthesis using IBMViaVoice and JavaSpeech API (Figure 3).

```
<?xml version='1.0' encoding='utf-8'?>
<configurationsequence nbrconfig="2">
        <configuration>
        <timecode>1</timecode>
        <body>body.gif</body>
        <head>head-front.gif</head>
        <eyes>eyes-open-happy.gif</eyes>
        <gaze>pupils-middle.gif</gaze>
        <facial>lips-open.gif</facial>
        <bothArms>null</bothArms>
        <leftArm>arm-left-hello1.gif</leftArm>
        <rightArm>arm-right-
hip.gif</rightArm>
        <speech>Hello, my name is LEA!
</speech>
        </configuration>

          <configuration>
          <timecode>8</timecode>
          <body>body.gif</body>
          <head>head-front.gif</head>
          <eyes>eyes-up-surprise.gif</eyes>
          <gaze>pupils-middle.gif</gaze>
          <facial>lips-down.gif</facial>
          <bothArms>null</bothArms>
          <leftArm>arm-left-hand-
down.gif</leftArm>
          <rightArm>arm-right-hand-
down.gif</rightArm>
          <speech>null</speech>
          </configuration>
</configurationsequence>
```

Table 2: Low-level specification of each modality in the LEA agent. Each configuration specification features the image to be displayed for each body part. The corresponding display is provided in Figure 2.

Figure 2: Screendump of the LEA agent corresponding to the specifications of Table 2.

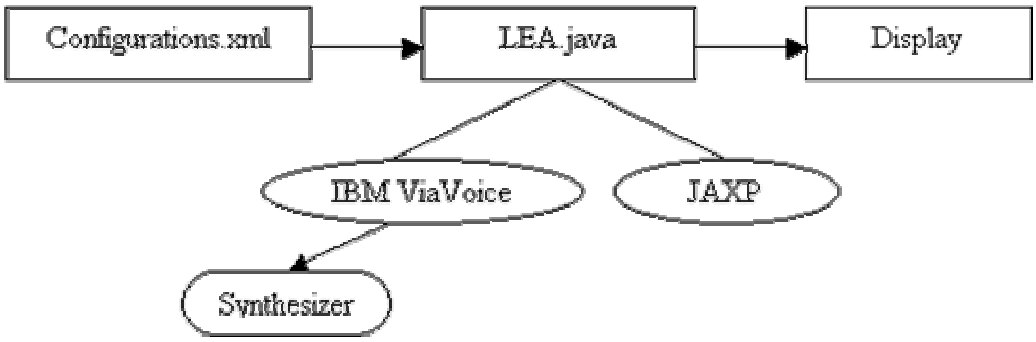

Figure 3: Current architecture of the LEA agent. The XML file containing a sequence of configuration specification is parsed by the LEA Java software with the JAXP API. Multimodal behavior is displayed via gif images and speech output using IBMViaVoice.

Towards "intermediate" specification of cooperation between modalities

We intend to augment the current specification of the LEA agent with "intermediate" level specification tags. This "intermediate" level of specification will be defined between the currently used low level of specification (ie. sequence of images) and a higher level of specification (ie. semantic representations, pragmatic and communicative goals…). This intermediate level of specification will be based on the Tycoon typology of cooperations between modalities (Martin et al. 2001). We believe that since this typology seems useful for the annotation of human multimodal behavior, it might also be useful to exhibit "natural" multimodal properties in agent behavior:

- ***Equivalence***: A cooperation by equivalence is defined by a set of modalities, a set of chunks of information, which can be displayed on either of the modalities and a criterion, which can be used by the agent to select one of the modalities. When several modalities cooperate by equivalence, this means that a chunk of information may be displayed as an alternative, by either of them.
- ***Redundancy***: Several modalities, a set of chunks of information and two functions define a cooperation by redundancy. The first function can be used to find out the common attributes in chunks to be presented by the different modalities, the second function is used as a fission criterion. If modalities cooperate by redundancy, this means that these modalities will present the same information (ie. the values of several attributes of displayed monomodal information will overlap).
- ***Complementarity***: Cooperation by complementarity is similar to cooperation by redundancy except that there are several non-common attributes between the chunks to be displayed by the different modalities.
- ***Specialization*** : Cooperation by specialization is defined by a modality, a set of modalities A and a set of chunks of information this modality is specialized in when compared to the modalities of the set A. When modalities cooperate by specialization, this means that a specific kind of information is always displayed by a single modality.

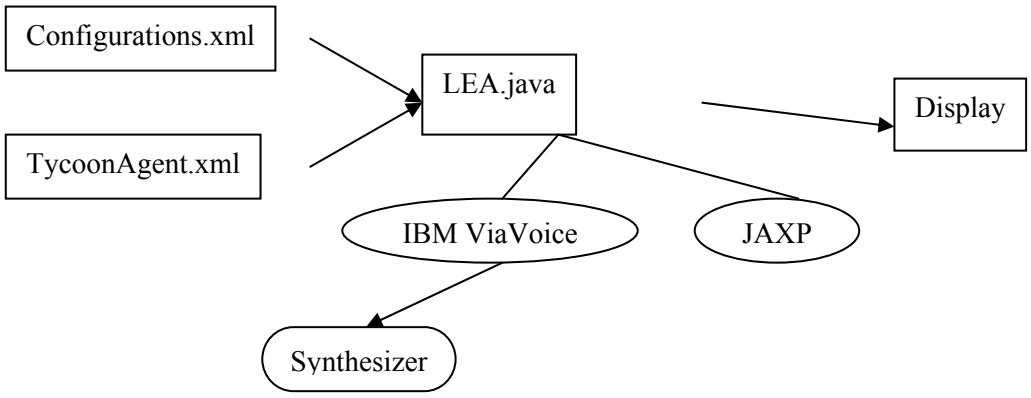

Figure 4: Independent specification of the agent's multimodal personality (TycoonAgent.xml) and the sequence of multimodal configurations (Configurations.xml).

Future directions

Such intermediate tags specifying cooperation between modalities might be integrated with the low-level tags in different ways.

One possibility is to use one file `TycoonAgent.xml` defining the multimodal behavior (or personality) of the agent thanks to Tycoon tags (ie. equivalence / redundancy…), and a second file `Configurations.xml` containing the initial presentation that the agent must achieve. The specifications provided in `TycoonAgent.xml` would then act as a filter of the presentation specified in `Configurations.xml` in order to extract the multimodal expressions of LEA (Figure 4).

Another possibility is to also include Tycoon tags in the configuration file itself. It would then be possible to make for example the agent more or less redundant at certain times. Example: <**redundancy**>right</**redundancy** > will lead the avatar to gesture with the hand, the body and gaze towards the right-hand side. "Cascaded multimodal style sheet" might be used: Tycoon tags provided in "Configurations.xml" would have priority. If there is not any, those of "TycoonAgent.xml" would be used as default multimodal behavior.

Future directions also include the use of multimodal input (speech recognition and 2D gesture) to interact with the agent.

Acknowledgments

The graphical design of the LEA agent has been achieved by Christophe RENDU who can be contacted at crendu@nexusanimation.com and +33.6.03.60.43.62

References

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H. and Yan, H. (1999). "Embodiment in Conversational Interfaces: Rea." Proceedings of the CHI'99 Conference, pp. 520-527. Pittsburgh, PA.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., Yan, H. (2001a). More than just a pretty face: conversational protocols and the affordances of embodiment. Knowledge-Based Systems, 14, 55-64.

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., Rich, C. (2001b) Non-Verbal Cues for Discourse Structure. Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics, pp. 106-115. July 17-19, Toulouse, France

Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjálmsson, H. (2002) MACK: Media lab Autonomous Conversational Kiosk. Proceedings of Imagina02. February 12-15, Monte Carlo.

Cassell, J., Vilhjálmsson, H., Bickmore, T.(2001c) BEAT: the Behavior Expression Animation Toolkit. Proceedings of SIGGRAPH '01, pp. 477-486. August 12-17, Los Angeles, CA.

Guerrin, F., Kamyab, K., Arafa, Y., Mamdani, E. (2001) Conversational Sales Assistants. Proceedings of the workshop on " Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents ", May 29, 2001, Montreal, in conjunction with The Fifth International Conference on Autonomous Agents. pp 35-40.

Gustavsson, C., Strindlund, L., Wiknertz, E., Beard, S., Huynh, Q., Marriott, A., Stallo, J.

(2001). Virtual Human Markup Language. http://www.vhml.org/.

Kipp, M. (2001a) Analyzing Individual Nonverbal Behavior for Synthetic Character Animation. In: C. Cave, I. Guaitella, S. Santi (eds.) Oralité et Gestualité - Actes du colloque ORAGE 2001, Paris: L'Harmattan, p 240-244.

Kipp, M. (2001b) Anvil - A Generic Annotation Tool for Multimodal Dialogue. Proceedings of Eurospeech 2001, pp. 1367-1370, Aalborg, September 2001.

Martin, J.C. & Kipp, M. (2002). Annotating and Measuring Multimodal Behaviour - Tycoon Metrics in the Anvil Tool. Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'2002), Las Palmas, Canary Islands, Spain, 29-31 may 2002

Martin, J.C. , Réty, J.H., Bensimon, N. (2002). Multimodal and Adaptative Pedagogical Resources. Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'2002), Las Palmas, Canary Islands, Spain, 29-31 may 2002 http://www.lrec-conf.org/lrec2002/index.html

Martin, JC., Grimard, S., Alexandri, K. (2001). On the annotation of the multimodal behavior and computation of cooperation between modalities. Proc. of the workshop on "Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents", May 29, 2001, Montréal, in conjunction with The 5th Int. Conf. on Autonomous Agents. pp 1-7.

Maybury, M. & Martin, J.-C. (2002) Proceedings of the workshop "Multimodal Resources and Multimodal Systems Evaluation". M. Maybury & J.-C. Martin (Eds). Third international conference on language resources and evaluation (LREC'2002), Las Palmas, Canary Islands, Spain, 1st June 2002

Silva, A., Vala, M., Paiva, A. (2001). The Storyteller : Building a Synthetic Character That Tells Stories. Proceedings of the workshop on "Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents", May 29, 2001, Montreal, in conjunction with The Fifth International Conference on Autonomous Agents. pp 53-58.

Traum, D., Rickel, J. (2001) Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. Proceedings of the workshop on " Representing, Annotating, and Evaluating Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents ", May 29, 2001, Montreal, in conjunction with The Fifth International Conference on Autonomous Agents. pp 27-33.