The Control of Agents' Expressivity in Interactive Drama

Nicolas Szilas[1,2] and Maurizio Mancini[1]

[1] LINC, IUT de Montreuil, 140 rue de la Nouvelle France,
93100 Montreuil, France
m.mancini@iut.univ-paris8.fr
[2] Department of Computing, Macquarie University,
NSW 2109, Australia
nicolas@ics.mq.edu.au

Abstract. This paper describes how conversational expressive agents can be used in the context of Interactive Drama. This integration requires some automatic tagging of the generated text, according to the current dramatic situation. Experimental results of the implemented prototype are presented.

1 Agents and Interactive Drama

1.1 Project Description

Interactive Drama (ID) is a new narrative form which takes place on a digital medium and where the audience plays an active role by deciding how to behave in the story, acting as a character. People are accustomed to "fake" ID with Interactive Fictions or adventure video games. In those cases, the freedom of the user as a character is very limited compared to the range of actions that characters are supposed to have in a story. True ID has not yet been reached but several prototypes are going in this direction [4], [7], [12], [14], [18].

Part of the difficulties of ID lies in the fact that it requires many diverse components. Furthermore, the research in ID is scattered and systems have been developed rather independently so far, with limited reuse of components developed externally. This paper aims at combining two components to create a more compelling ID: an expressive Embodied Conversational Agent (ECA) system called Greta [9] and an interactive narrative engine called IDtension [14]. The primary goal of this integration is to convey the story with emotion through using emotionally expressive agents.

Even if most narrative models focus on the cognitive dimension of storytelling, emotion is now recognized as a key feature of narrative and drama. Regarding the narrative film in particular, the narrative structure itself is constructed in order to produce some emotions such as hope or fear [16]. According to Noel Carroll, emotions even constitute the condition of film understanding. They allow the viewer to focus his/her attention on important elements in the narrative [3].

We assume that these general data about emotion in narrative still hold in the interactive case. ID should trigger the user's emotions and this paper investigate one of the main sources of emotion in Drama: the emotions expressed by the characters.

In fact, emotional agents have been the basis of ID since the birth of the field [6], but our project differs from previous systems for the following reasons:

- The story is dynamically generated by a narrative engine rather than by a simulation of agents [15]. Thus, we raise the theoretical and practical issue of how to coordinate emotions at both narrative and behavioral levels;
- The story events themselves are generated (including the texts that are used in dialogs), which prevent any scripted solution for emotion generation;
- The expressive agent component enables the system to generate precise and rather realistic facial expressions, which opens the way to more engaging ID.

1.2 Generative Interactive Drama: IDtension

IDtension is a long term research project, started in 1999, which aims at solving the paradox of interactive narrative (see [14], [15] for details). It is built around a narrative engine which dynamically generates a story and lets the user choose the actions of one character. It operates at the logical level, in a medium independent manner. It contains the following four main distinctive features:

1. The story model is a fine grain model, in the sense that it manipulates elementary actions rather than larger units like for instance beats [7] or scenes. Such a fine grain model provides more interactivity but prevents the author to easily craft each scene. These scenes are generated by the system.

2. The system includes a model of the user aimed at estimating the impact of each possible action on the user according to several narrative criteria [14]. Some of these criteria focus on believability. Other criteria such as *complexity* and *conflict* are only guided by narrative concerns. Some of them are related to the user's emotions: the system is able to trigger one action instead of another with the intention to trigger a given emotion in the user. Conflict for example occurs when the user's character has a specific goal while the mean to reach that goal (a *task*) goes against his own ethical values. Such conflict is the core of classical drama and produce strong emotional reaction in the viewer.

3. The articulation of actions is twofold. IDtension considers generic actions and specific tasks. Generic actions stem from narratology [2], [17]. They are, for instance, inform, encourage/dissuade, accept/refuse, perform, felicitate/condemn. Tasks are specific to a story: kiss, hug, slap (in a romance story) or threaten, torture, kill (in a *roman noir)*, etc. This makes it possible to handle complex actions like “John tells Mary that Bill has robbed her jewels” without requiring from the author to explicitly enter them into the system. The narrative engine handles logical forms such as *Inform(John, Mary, have_finished(Bill, rob , [jewels,Mary])*. The author only specifies the task (rob), the characters (Mary, John, Bill) and the objects (the jewels) in the story.

4. IDtension explicitly processes the notion of (ethical) values. Values are thematic axes according to which each task is evaluated. Such values include honesty, friendship, family, etc. This mechanism adds beyond the pure performative dimension, another dimension to the story namely the axiological dimension. The *Model of the User* processes those values to evaluate some narrative criteria, in particular *conflict*.

Action selection is performed in three steps:

1. The *Narrative Logic* generates the set of all possible actions at a given time in the narrative via a set of narrative rules.
2. The *Model of the User* assesses each of these actions according to its estimated impact on the user. The actions can then be ranked.
3. The *Theatre* displays the selected action by generating the text with a template based approach and by sending it to the virtual characters.

The system alternates actions chosen by the engine and actions chosen by the user.

1.3 Expressive Embodied Conversational Agents: GRETA

ECAs are entities with a human-like appearance capable of taking part in dialogs with the users. They communicate multimodally with their human interlocutors (or with other agents) through voice, facial expression, gaze, gesture, and body movement.

Fig. 1. The Greta agent

Greta (see Fig. 1) is an ECA system that given an XML-tagged text as input provides some animation files that drive an embodied representation of a virtual human as described in [9]. The input file contains both the text that the agent has to speak and the information on the communicative intent of the conversation, defined using an XML-based language called APML (Affective Presentation Markup Language). APML is based on the taxonomy of communicative functions defined in [11] by Poggi. The tags of APML provide information on the Speaker's Mind - on his/her beliefs, goals and emotions:

– belief tags — they convey information about the certainty/uncertainty of the beliefs that the Speaker is referring to and their source (for example, long-term memory).
– goal tags — they refer to the Speaker's goals, for example the generic goals of giving general information. Or they can help the Speaker underline the most important part of the sentence, specify logical relation between different parts of the discourse, manage the evolution of the conversation (for example turn-taking).
– emotion tags — they are meant to convey the affective states of the Speaker that will mainly influence the facial expression and body gestures that the agent will choose to show.

In the example displayed in Table 1, we can distinguish the *performative* tags on lines 2 and 17 that represent the goals the agent wants to achieve with the two sentences inside of them. On line 9 and 19 the presence of the *rheme* tags means that the surrounded text is the important part of the discourse. Finally, some emotional content ("resentment") is specified on line 19.

Moreover, APML provides some special tags and attributes to drive the speech synthesizing process. For this purpose, the tag *boundary* (lines 14, 25) and the attribute *x-pitchaccent* (lines 4, 11, 21) have been introduced, as explained in [13]:

- boundaries : they are perceived as prosodic phrase-final rising pitch and/or lengthening.
- x-pitchaccents : they are perceived as word-level stress, accent or emphasis.

Table 1. A typical APML input file for the Greta agent

```
1. <apml>
2. <performative type="warn">
3. <theme>
4. <emphasis x-pitchaccent="Hstar">
5. Please
6. </emphasis>
7. do not tell me what works
8. </theme>
9. <rheme>
10.for
11.<emphasis x-pitchaccent="Hstar"
   deictic="self">
12.me
13.</emphasis>
14.<boundary type="LL"/>
15.</rheme>
16.</performative>
17.<performative type="suggest">
18.Would you just please
19.<rheme affect="resentment">
20.mind your
21.<emphasis x-
   pitchaccent="Hstar"
   adjectival="foreign">
22.own
23.</emphasis>
24.business
25.<boundary type="H"/>
26.</rheme>
27.</performative>
28.</apml>
```

Given an APML file, the Greta system looks in a dictionary of signals for the gestures/facial expressions associated to the communicative functions specified by the tags in the input file. A gesture/face planner is responsible for instantiating these signals according to temporal and contextual constraints (e.g. coarticulation between gestures/facial expressions). *Festival* (Univ. of Edinburgh) is used for speech synthesis.

The exact procedure the agent performs in generating animation is:

1. Input — the starting point is the APML file with tags representing the communicative functions listed above (beliefs, goals, emotions).
2. Signals selection — the system looks for the associated gesture/facial expressions to each tag in the signal dictionary.
3. Synchronization — the signals are instantiated according to temporal and contextual constraints (e.g. coarticulation). Each tag of the APML file corresponds to some text (the text which is "nested" in the tag). So from the duration of the speech associated to the text tags, the exact starting and ending times of each tag (and consequently for each signal) are calculated. At this point the correct temporal characteristics of the facial expressions [1] and gestures [5] are generated.

2 Theoretical approach

2.1 A communication-based model for expressivity

The relationship between agent's expressivity and narrative expressivity is as complex as the relationship between character and narrative.

One should be careful not to reduce a narrative to the simulation of a fictional world. A narrative contains only a limited number of actions usually discontinuated simply evoking a fictional world in the receiver but does not contain a simulated world. In addition these limited actions are carefully crafted to convey a certain meaning, to trigger certain emotions, which goes beyond a world simulation. Hence the famous Hitchock's quote: “Drama is life with the dull bits cut out”.

Research in embodied agents is usually performed (and funded) in a context of a world simulation, in order to simulate humans (“virtual humans”, “lifelike characters”,

etc.). Research in emotional agents in particular aims at providing mechanisms for an agent to react properly to given situations by expressive realistic emotions. “Believable agents” [6] aim at refocusing the simulation towards its expressivity. However, the actions of believable agents are still resulting from the simulation of humans in the environment even if those mechanisms tend to be more expressive (e.g. exagerated).

According to the general view of drama introduced above (drama is more and less than a world simulation), we meet two sets of problems in directly applying virtual human research to ID:

- Simulating a set of emotional believable agents and then selecting only the parts that are shown to the user seems to constitute quite a detour. This simulation is a challenge in itself.
- Integrating expressive features calculated by the narrative engine with those calculated according to the world simulation is problematic.

For those reasons we decided to focus only on what is perceived by the user, namely the actions on the screen. The actions calculated by the narrative engine come with the necessary information to calculate the corresponding expressive features. In particular, those actions are calculated according to a goal structure, which constitutes a key feature in emotions [8].

We call this approach a communication-based model, because it links emotions to actions that are conveyed to the user rather than calculating it according to the agent's states. Between two actions, no emotion is calculated. This previous affirmation is hard to follow in a virtual environment because characters are visible between narrative actions. The solution is to assign to an agent the emotion of the previous narrative event it was involved in with a decreased intensity in time.

2.2 Authoring expressivity

In classical (non interactive) drama, the expressivity of a given sentence is produced by the actor based on the annotated script and with the help of the director. In the digital context such human skill is not available during the performance. To compensate two actions are made available:

- manually annotating each sentence in detail, so that the agent “only” follows gestural and prosodic directions;
- calculating on the fly these expressive directions according to a set of predefined linguistic rules and data.

In the context of generative Interactive Drama, a concept such as “each sentence” is nonexistent. These sentences are generated on the fly, and even if their number is finite, it would be much too large to allow any manual annotation. Only templates are available (sentences with gaps to fill), which makes expressivity authoring uneasy, because each piece of text could appear in different contexts. Moreover, this kind of annotation using the APML language requires a high level of expertise in linguistics as well as specific training [13].

However, relying solely on automatic tagging of text does not provide the best quality of expressivity. Doing so would furthermore frustrate the author who would have no control over agent's expressivity.

Thus, a hybrid approach is suggested. The system automatically creates expressive annotations of the text, while the author is able modify to a certain extent the expres-

sive rendering. The author's work would require some easy tagging of the text, which will serve as a basis for APML generation. This strategy has been partially implemented, as described in Section 3.4.

3 Implementation

3.1 Technical integration

Some modifications of the Greta system have been made, in order to integrate it with the IDtension system. A TCP/IP server able to accept external connections has been added. It has been implemented as a parallel thread that constantly waits for incoming TCP/IP connections from any external source (client). In this manner another process (that is, the IDtension system) can communicate to Greta and send data as text strings.

The IDtension system decides which APML data has to be transmitted by taking the logical form of an action and generating the corresponding text to be spoken with the tags, according to the rules described below. It then connects to the Greta's server and sends the APML data through the TCP socket as a sequence of strings. Each time the sending of APML data is completed, the Greta system elaborates it as described in section 1.3 and generates the corresponding graphical animation.

3.2 Rules for the tagging of emotions

Two types of emotions are handled in the narrative engine: emotions that are associated to the user and managed in the *Model of the User* and emotions that are associated to the characters and managed by the *Theatre*. Only this second type of emotions will be considered. In addition, only emotions associated with the character's current action are considered, not those triggered in reaction to other character's actions.

Table 2. Emotions implemented in the system.

Type of emotion	Emotion	Types of action involved
before task performance	enthusiasm	inform {want, have_begun}
	worry	inform {want, have_begun, hinder}; accept
after task performance	satisfaction	inform {have_finished}
	disappointment	inform {have_been_blocked}
after task performance, value-related	disgust	inform {can,want, have_begun}
	anger	dissuade, condemn
	shame	inform {have_begun, have_finished}

The method of emotion association is related to the cognitive structure of emotions described in [8], but instead of identifying emotional states of the agents, the system focuses on actions, as suggested in [10]. Seven emotions have been implemented so far (see Table 2).

Each automatic emotional tagging is defined by the following elements:

- types of actions the emotion can be associated to,
- condition of triggering the emotion,
- calculus of the intensity.

Two of the these emotions will be detailed here: worry and disappointment.

Worry: this emotion occurs when the character perceives that there is a chance of failure of the agent's performance.

The actions for which worry could be expressed are all actions related to a task before it is accomplished (see Table 1).

The "worry" emotion is triggered if all the following conditions are met:

- The importance of the goal for the character is superior to a given threshold.
- The perceived risk cumulated over all known obstacles for the corresponding task is superior to a given threshold.

Note that these data are available in the narrative model [14].

The intensity is calculated differently according to the action:

- if the action is an "inform hinder", which means that the character is talking about a specific obstacle on the task:

$$\text{intensity} = \text{importance}(\text{goal}) \times (1\text{-perceived\_risk}(\text{obstacle}))$$

- in all other cases:

$$\text{intensity} = \text{importance}(\text{goal}) \times \Pi_i\,(1\text{-perceived\_risk}(\text{obstacle}_i))$$

where *i* ranges in all known obstacles.

Disappointment: this emotion occurs when an obstacle has been met.

It is expressed with the action "inform have_been_blocked", when a character informs another that s/he met an obstacle during the execution of a task.

The "disappointment" emotion is triggered if all the following conditions are met:

- The importance of the related goal for the agent is superior to a given threshold.
- The obstacle was met recently, that is the difference between the current date and the date of the triggering of the obstacle is below a given threshold.

The intensity equals the importance of the goal for the character.

3.3 Rules for the tagging of boundaries

Adding the boundary tags requires three steps:

- segment the text,
- choose between theme and rheme,
- choose between the different types of boundaries.

These steps when performed by a human are quite difficult. They require some knowledge about theories on English prosody as well as practice [13]. Therefore, the rules proposed below are a strong approximation of the proper coding. We intend to improve these rules in the future.

To segment the text, two rules are used:

- Each new predicate corresponds to a new segment. For example, the action *inform(Joe, Mary, want(Joe, swim))*. ("Joe tells Mary that he wants to swim") is broken up into two segments: "Joe tells Mary that" and "he wants to swim".
- Each punctuation marks a new segment.

The choice between *theme* and *rheme* is made based on the two following rules:

- The larger segment of a sentence which correspond to a single action is a rheme (excluding information).
- For information, the content of the information is a rheme (*what* is informed) and the rest is a theme. This corresponds to the fact that what is new in a given information is the content, not the surrounding text. For example, in the sentence: "You know what? Joe bought a car", the segment "You know What?" is the theme, while the segment "Joe bought a car" is the rheme (new information).

The choice of the type of boundaries has been limited to two types of boundaries: LH (rising pitch) and LL (falling pitch). The following rules are applied [13]:

- For boundaries related to a predicate, if it is a theme the type "LH" is given, and if it is a rheme, the type "LL" is given.
- For boundaries related to the punctuation, the type is "LL" if it is a "period" and "LH" for all other punctuation marks.

3.4 Rules for the tagging of emphases

The tagging of emphases is performed in two steps:

- detection of logical components of the action that should be emphasized,
- choice of a method for emphasizing.

The components to which an emphasis can possibly be applied are characters, objects, places, goals, tasks and obstacles. The decision whether an emphasis should be applied or not is related to the need to express some contrast on this component [13]. A component is contrastive when it brings some new information to the addressee. For example, consider the following dialog:

Joe to Mary: "I want to swim"

Mary to Joe: "If you swim, be careful then, it's dangerous"

In the first line, the new information is "swim", but in the second line, the new information is "it's dangerous".

The narrative engine manages contextual information in terms of narrative sequences or processes [14]: to most actions is associated an "initiating action" to which the current action is responding (which is not necessary the action played previously).

Once the decision to add an emphasis to a component has been made, the next step is to surround the corresponding text with the emphasis tags. However, for complex components putting stress on a long piece of text is not adviced. For example, in the sentence "to evacuate the building, you should trigger the fire alarm" the emphasis should be put on "fire alarm" not on "trigger the fire alarm". This choice is difficult to make automatically because it requires either some knowledge about the real world ("fire" and "alarm" are dramatic elements) or linguistic knowledge ("trigger" is a verb and "the" an article, while "fire alarm" is a substantive).

We have chosen to solve this issue by letting the author define which words in the text s/he writes should be emphasized. In the template for the task *triggerFireAlarm*, the author puts "trigger the *fire alarm*". Finally, the rule is as follow: if the component contains an author-defined emphasis, then only the parts defined by the author should be emphasized.

4 Results

An interactive scenario named “Mutiny” has been implemented, which is about four prisoners trying to escape from an old galleon by starting a riot. The user plays "Jack", one of the prisoners. S/he interacts by selecting an action into a large list. Actions are displayed with both text and Greta's head (body movement is also automatically generated by Greta, but is not displayed). Fig. 2 represents a possible path that displays four different emotions. Boundaries and emphases are not represented but can be heard in the synthesized speech and seen in the animation's dynamics.

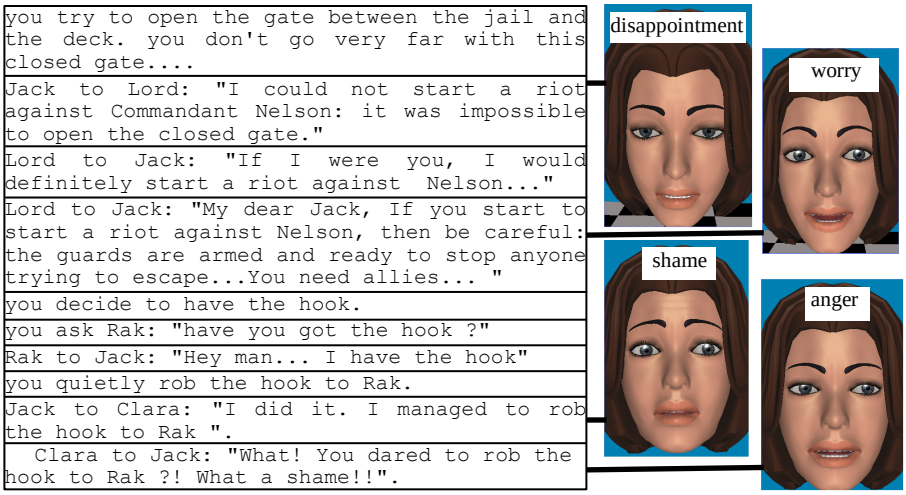

Fig. 2. Text and screenshots of one specific interactive session, highlighting four facial expressions. Note that only one face model is used, even if different characters are involved (Jack for disappointment and shame, Lord for worry and Clara for anger). The emotions of shame and anger are related to characters' ethical values (see Section 1.2).

5 Conclusion

A system which integrates a highly generative ID engine (IDtension) with an ECA (Greta) has been described. This opens the way to more engaging ID where characters' facial and body animations are not pre-scripted but generated based on the current dramatic situation. To reach that goal, we have adopted a communication-based approach in order to focus on action, the core of drama.

As an extension of this research, we intend to study in detail the relationship between the emotions displayed by the agents and to the emotions triggered in the user.

Thus, the agent's expressivity will be generated not only by their in-fiction behaviour but also by the narrative effects calculated by the narrative engine. Such a study would benefit from the data available from Film Studies on how the audience feels empathy or counter-empathy towards the characters in a movie.

Acknowledgments

This research is supported by the *Maison des Sciences de l'Homme de Paris Nord* and by the *Australian Research Council* (Linkage Grant #LX560117).

References

1. Bevacqua, E., Mancini, M., Pelachaud, C.: Speaking with Emotions AISB 2004 Convention: Motion, Emotion and Cognition, University of Leeds, UK, (2004)
2. Bremond, C.: Logique du récit. Seuil, Paris (1974)
3. Carroll, N.: Beyond Aesthetics. Cambridge University Press, Cambridge (2001)
4. Crawford, C: Assumptions underlying the Erasmatron interactive storytelling engine. In Papers from the AAAI Fall Symposium on Narrative Intelligence, Technical Report FS-99-01. AAAI Press, Menlo Park (1999)
5. Hartmann, B., Mancini, M., Pelachaud, C.: Formational parameters and adaptive prototype instantiation for mpeg-4 compliant gesture synthesis. In Proceedings of the Computer Animation 2002. IEEE Computer Society (2002) 111
6. Kelso, M., T., Weyhrauch, P., Bates, J.: Dramatic Presence. TR CMU-CS-92-195, Carnegie Mellon University, Pittsburgh (1992)
7. Mateas, M., and Stern, A.: Towards Integrating Plots and Characters for Interactive Drama. In Dautenhahn K. et al. (eds): Socially Intelligent Agents. Kluwer Academic Publishers, Norwell, Dordrecht (2002) 221-228
8. Ortony A., Clore G., L., Collins A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1988)
9. Pelachaud, C., Bilvi, M.: Computational model of believable conversational agents. In Huget, M.P., ed.: Communication in Multiagent Systems. Volume 2650 of Lecture Notes in Computer Science. Springer-Verlag (2003) 300–317
10. Poggi, I.: Emotions.Un modèle en termes de buts et de croyances. Unpublished seminar, Montreuil, France, July, 12, 2004 (2004)
11. Poggi, I.: Mind markers. In M. Rector, I. Poggi, N.T., ed.: Gestures. Meaning and use. University Fernando Pessoa Press, Oporto, Portugal (2003)
12. Sgouros, N. M.: Dynamic Generation, Management and Resolution of Interactive Plots. Artificial Intelligence, 107,1 (1999) 29-62
13. Steedman, M.: Using APML to Specify Intonation. Unpublished Tutorial Paper (2004). ftp://ftp.cogsci.ed.ac.uk/pub/steedman/apml/howto.pdf
14. Szilas, N.: IDtension: a narrative engine for Interactive Drama. In Göbel et al. (ed): Proc. TIDSE'03. Frauenhofer IRB Verlag (2003)
15. Szilas, N.: Interactive Drama on Computer: Beyond Linear Narrative. In Papers from the AAAI Fall Symposium on Narrative Intelligence, Technical Report FS-99-01. AAAI Press, Menlo Park (1999) 150-156
16. Tan, E.: Emotion and the structure of narrative film. Film as an emotion machine. Erlbaum, Mahwah (NJ) (1996)
17. Todorov, T. Les transformations narratives. Poétiques, 3 (1970) 322-333
18. Young, R. M., Riedl, M. O., Branly, M., Jhala, A., Martin, R.J. Saretto, C. J.: An architecture for integrating plan-based behavior generation with interactive game environments, Journal of Game Development, 1, 1 (2004)