Papous: The Virtual Storyteller

André Silva, Marco Vala, and Ana Paiva

IST / INESC-ID, Rua Alves Redol 9, 1000-029 Lisboa, Portugal
andre.silva@gaips.inesc.pt, marco.vala@gaips.inesc.pt,
ana.paiva@inesc.pt

Abstract. This paper describes the development of Papous, a Virtual Storyteller. Our ultimate goal is to obtain a synthetic character that tells stories in an expressive and believable way, just as a real human storyteller would do. In this paper we describe the first version of Papous, our virtual storyteller. Papous can be seen as a virtual narrator who reads a text enriched with control tags. These tags allow the storywriter to script the behaviour of Papous. There are four types of tags: behaviour tags, where a specific action or gesture is scripted; scene tags, that allows for Papous to change the scene where he tells the story; illumination tags, to allow a new illumination pattern of the scene; and emotion tags, to change the emotional state of Papous. The texts, enriched with these tags, are then processed by Papous' different modules, which contain an affective speech module and an affective body expression module. In this paper we will provide details of the speech, gestures and environment control actions taken by each of the modules of Papous architecture.

1. Introduction

Stories and storytelling are a constant presence in our lives since early childhood. Children like to be told the same story, over and over again, without getting tired of the exact same words. And, the storyteller plays a fundamental role in children's stories. In fact, a storyteller can turn a story into a good or a bad one. A good storyteller is able to drag us into the story, keep our attention and free our imagination. The use of the voice, facial expressions, and the appropriate gestures, are basic ingredients for transforming the content of a simple story into the most fascinating narrative we have ever heard.

But this need for a storyteller to be expressive, to balance the words and the tone, to use gestures appropriately, etc, poses major research challenges if one aims at building a "synthetic" storyteller. However, recent developments of embodied agents [Badler et al. 2000, Cassell 2000, Cassell et al. 1999, Cassell et al. 2000, Churchill et al.] have, during the last few years, shown amazing advances.

Thus, aiming at a synthetic storyteller, we created Papous. The ultimate goal is for Papous to be able to tell the content of a story in a natural way, expressing the proper emotional state as the story progresses and capture the user's attention in the same way a human storyteller would.

A. de Antonio, R. Aylett, and D. Ballin (Eds.): IVA 2001, LNAI 2190, pp. 171-180, 2001.

At the present time, our work is still in an early stage. Papous simply acts as a virtual narrator who reads a text enriched with control tags. Such tags allow the storywriter to control the character (its actions and/or emotional state) and the surrounding environment (to achieve a correlation between the story and the ambience). This approach is similar to the one taken by Allison Druin [Druin et al. 1999] where children annotated text so that a robot could produce the appropriate emotions when the story was narrated.

In this paper we will describe the current architecture of Papous, the tags associated to a text, and how storytelling is performed taking into account the described architecture. We will provide details of the speech, gestures and environment control actions taken by each of the modules of Papous architecture. Finally we will describe some results and future work.

2. Architecture

The architecture of Papous, as depicted in Fig. 1, has five components: the *Input Manager*, the *Environment Control*, the *Deliberative Module*, the *Affective Speech* and the *Affective Body Expression*.

The *Input Manager* is the component responsible for processing the text file that contains the story, checking it for syntax and semantic errors, and taking the necessary actions to ensure that the data is correct and ready for the other components to process.

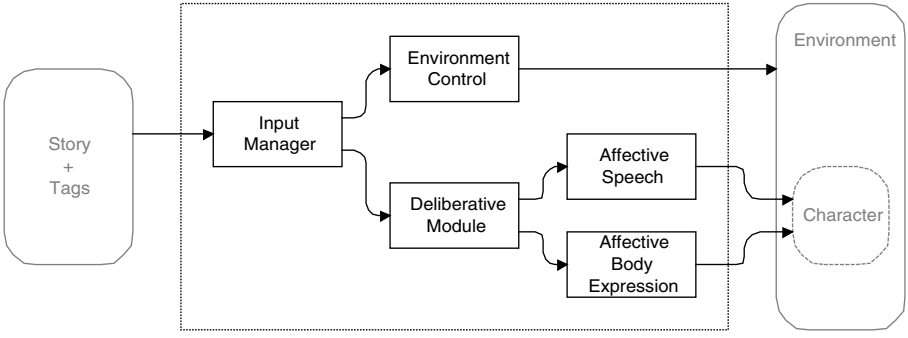

Fig. 1. Architecture

The *Environment Control* is responsible for managing the environment where the character is immersed.

The *Deliberative Module* acts as the mind of the character and, therefore, contains its internal emotional state and is responsible for controlling the character's actions.

The *Affective Speech* is responsible for the voice of the character.

The *Affective Body Expression* is responsible for the appearance of the character.

3. Input Manager

The *Input Manager* receives as input the annotated story file and a set of configuration files. Table 1 summarizes the four types of tags available and explains the function of each one.

Table 1. Tag types

Tag Type	Function
Behaviour (1)	Indicate an action that the character should perform (e.g. **<1*big>**)
Scene (2)	Specify a new scene where the character should be integrated (e.g. **<2*house>**)
Illumination (3)	Specify a new illumination pattern (e.g. **<3*day>**)
Emotion (4)	Explicitly modify the emotional state of the character (e.g. **<4*happiness*80>**)

The list of available tags of each type is defined in a configuration file and depends solely on the available scenes and animations.

We have defined a very small set of tags for demonstration purposes. Fig. 2 illustrates a possible use of these tags.

<2*house> **<4*happiness*80>** Hello everybody! I am extremely happy today! Lets take the usual tour, ok? **<4*happiness*50>** This is the house I live in. It is a very **<1*big>**big house**<~1*big>**.
Want to go outside? **<2*street>** Ahhh.... isn't this nice? I live in a **<1*small>**small town**<~1*small>** right by the sea... Hmmm....I think it will be dark soon... **<3*night>** **<4*fear*80>**Oh, I'm so afraid of the dark... Maybe we should get back in the house, right?
<2*house> Hey! Do you like stories? I bet you do! You know, I have a friend named Alex. He is a writer, and he is **<1*tall>**very tall **<~1*tall>**. Much taller than me!... **<4*happiness*20>** I haven't seen him in a while... and that makes me kind of sad...
<4*happiness*50> Anyway, I also have **<1*short>**a very short friend **<~1*short>** named Paul. Hey! We have been talking for a long time... It is almost morning! **<3*day>**
<4*surprise*90>What a marvellous day! **<4*happiness*50>** Come back soon, ok? Bye,bye...

Fig. 2. Example of annotated text

The storywriter is free to use the tags as he pleases, but he should take in consideration the context of the story. For example, if the writer wants to emphasise a particularly scary part of the story, he should specify the appropriate emotional state.

The chosen emotional state will change the voice and the behaviour of the character and, therefore, suit the writer's intentions.

The *Input Manager* parses the annotated text and generates tag-oriented information that is sent to the *Environment Control* and *Deliberative Module* components.

4. Environment Control

The *Environment Control* component receives *scene* and *illumination* tags and changes the environment accordingly.

For demonstration purposes, we have chosen two scenes to immerse the character that can be toggled at anytime during story time. These scenes are briefly described in Table 2.

Table 2. The chosen scenes

Scene	Description
House	This scene represents the inside of a normal house, with some furniture and a fireplace. Decorative paintings exist in the wall, as well as a window showing the outside.
Street	This scene represents a city street. An old building, an alley with a garbage container, two streetlights and a garage.

The same way, we created two illumination patterns that can be applied to both scenes, although the lights that define these two patterns may differ from one scene to the other. Table 3 briefly describes the implemented illumination patterns.

Table 3. The illumination patterns

Illumination pattern	Description
Day	Abundant and natural light. Blue sky can be seen. This is the default illumination pattern.
Night	Less light. Sunset sky can be seen. This illumination pattern includes the use of spotlights.

To achieve smooth transitions between different scenes and illumination patterns we use a *fader* that acts like a theatre curtain. In its normal state, the *fader* is transparent and the scene is completely visible. When there is the need to perform a scene exchange (or an illumination pattern exchange), the visualization window starts to fade out, becoming black and hiding the scene. At this time, the scene and / or the illumination pattern is exchanged, and when it is ready, the visualization window fades in, becoming transparent, and allowing the user to view the scene again.

New scenes and illumination patterns could be added to the application, creating a series of possible environments in which to immerse the character.

5. Deliberative Module

The *Deliberative Module* receives *emotion* and *behaviour* tags and sends commands to the *Affective Speech* and the *Affective Body Expression* components.

The *emotion* tags update the internal emotional state indicating which emotion should be changed and the new value that it must have. Internally, the emotional state of the character is represented by a set of numerical discrete values. Figure 3 indicates the thresholds established for the three emotions used for demonstration purposes.

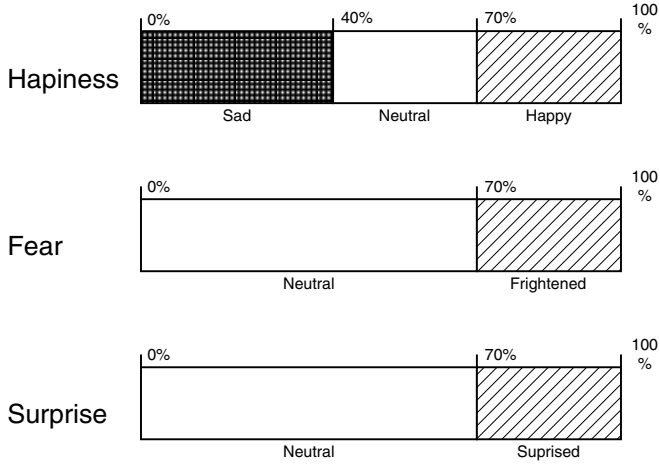

Fig. 3. Emotion thresholds

The emotional state affects the voice and the behaviour of the character. Of course, other emotions and thresholds could be defined, and further divisions could be considered to provide a richer control over the character.

The *behaviour* tags are associated to explicit gestures and result in direct commands to the *Affective Body Expression* component.

6. Affective Speech

The *Affective Speech* component receives sentences and the current emotional state from the *Deliberative Module*, and synthesizes the sentences using the voice to express the current emotions.

The precision with which we control the character's voice depends mostly on the underlying text-to-speech (TTS) system. We used a TTS system that allows the control of seven parameters to completely define the voice. These parameters are explained in Table 4.

Table 4. Voice parameters

Parameter	Description
Pitch baseline	Controls the overall pitch of the voice; high pitches are associated with women, and low pitches with men.
Head size	Controls the deepness of the voice.
Roughness	Controls the roughness of the voice.
Breathiness	Controls the breathiness of the voice; the maximum value yields a whisper.
Pitch fluctuation	Controls the degree of fluctuation of the voice.
Speed	Controls the number of words spoken per minute.
Volume	Controls the volume of the voice, i.e., how loud it sounds.

To transmit emotions through the voice we established a series of relations between emotions and voice parameters based in theories of the interrelationship between speech and emotion [Scherer 2000]. Table 5 indicates which parameters should be changed in order to transmit the emotion we intend through the character's voice.

Table 5. Emotions / TTS parameter correlation

Emotion	Parameter	Action
Happiness/Sadness	Speed	Increase/Decrease
	Pitch Baseline	Increase/Decrease
	Pitch Fluctuation	Increase/Decrease
Fear	Pitch Baseline	Decrease
	Pitch Fluctuation	Increase
	Breathiness	Increase
Surprise	Pitch Baseline	Increase
	Pitch Fluctuation	Increase
	Speed	Decrease

Another important aspect is the correct use of pauses, since pauses are very important to achieve natural speech. It is of critical importance that the voice pauses appropriately, considering the punctuation marks used.

To achieve the correct treatment of text pauses we classify the text into categories, which allow the application to be more specific in processing and interpreting the different punctuations marks that appear in the input file. Table 6 explains the different categories and the pause length associated with each of them.

Table 6. Text pauses

Text Category	Form*	Pause (ms)
Words	<Sentence>	100
Exclamation	<Sentence> !	500
Interrogation	<Sentence> ?	200
Period	<Sentence> .	200
Comma	<Sentence> ,	150
Omission points	<Sentence> ...	800

* <Sentence> = [a-zA-Z\']+

7. Affective Body Expression

The *Affective Body Expression* component receives the current emotional state from the *Deliberative Module* and changes the character body in order to express the desired emotions. It can also receive commands to perform gestures explicitly indicated in the story (using *behaviour* tags).

We use the body expression component provided by the SAFIRA toolkit [André et al. 2001]. This component is able to perform real-time blending between animations and body postures to convey the desired emotions.

However, at the current state of development, the emotions affect only the face of the character. For demonstration purposes we considered two facial animations (happy and sad) that are related with the *happiness* threshold.

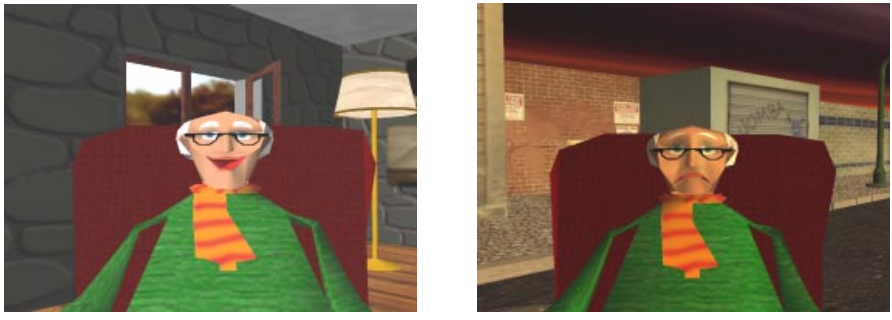

Fig. 4. The character is *happy* (left) or *sad* (right)

We have also defined a set of iconic gestures (*big*, *small*, *tall* and *short*) that can be explicitly indicated in the story. The writer should be careful in using *behaviour* tags to perform explicit gestures, as they only benefit the story if the performed action is coherent with the current story context.

Fig. 5. The character is indicating something is *big* (left) or *small* (right)

Fig. 6. The character is indicating someone is *tall* (left) or *short* (right)

To hide the absence of lip synchronization we use a single gesture that sequentially opens and closes the Papous' mouth. Since the speech is combined with the other gestures (specially the iconic gestures), the problem of lack of lip synch is reduced.

8. Concluding Remarks

The overall goal was achieved and Papous can already convey some emotions and tell a story that is amusing and a delight to hear. So, in general, the overall approach (architecture and module design) seems to be adequate for the intended purpose.

However, Papous has some limitations that must be improved. The most noticeable is the TTS system, which does not provide a great deal of flexibility when it comes to using its parameters to express the emotions we want. The voice seems to be more synthetic than we had hoped for.

The bodily expression is understandable, but limited by the number of available animations. From time to time, the absence of lip synchronization is also very noticeable.

Note that the current state of development of the project does not introduce a great degree of autonomy. In reality, the character is explicitly and externally controlled by

the input file, which works almost as a scripting language. Naturally, the character's autonomy and sensitivity to context will be further developed, as an obvious evolution for the storytelling character.

9. Future Work

The aspect we intend to improve the most is the autonomy of the storytelling character. The idea is to automatically detect the emotional, behavioural and environmental changes from the text, without using tags. Further, Papous will try to get some input from the user and environment (user's reaction to the story) and adapt some parts of the storytelling to that user.

We also intend to replace the TTS system with one capable of guaranteeing an affective speech system as proposed by Cahn [Cahn 1990] (with the one provided by the SAFIRA toolkit [André et al. 2001]).

Further, we will enlarge the available database with new scenes and animations to enhance the semantic richness of the application.

We are also considering embedding Papous into applications such as Teatrix [Paiva et al. 2001]. To do that, the text needs to be generated automatically from within the application.

Acknowledgements. The authors would like to thank Marco Costa and Fernando Rebelo for the artwork, and André Vieira, Filipe Dias, José Rodrigues and Bruno Araújo for their help, ideas and comments.

This work has been partially supported by the EU funded SAFIRA project number IST-1999-11683. Ana Paiva has also been partially supported by the POSI programme (do Quadro Comunitário de Apoio III).

References

[André et al. 2001] André E., Arafa Y., Gebhard P., Geng W., Kulessa T., Martinho C., Paiva. A., Sengers P., and Vala M.: SAFIRA Deliverable 5.1 – Specification of Shell for Emotional Expression (2001). http://gaips.inesc.pt/safira

[Badler et al. 2000] Badler N.I., Zhao L., and Noma T.: Design of a Virtual Human Presenter (2000).

[Cahn 1990] Cahn, J.E.: The Generation of Affect in Synthesized Speech. Journal of the American Voice I/O Society, Vol. 8 (1990) 1-19.

[Cassell 2000] Cassell J.: Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents (2000).

[Cassell et al. 1999] Cassell J., Bickmore T., Billinghurst M., Campbell L., Chang K., Vilhjálmsson H., and Yan H.: Embodiment in Conversational Interfaces: REA (1999).

[Cassell et al. 2000] Cassell J., Bickmore T., Campbell L., Vilhjálmsson H., and Yan H.: Human Conversation as a System Framework: Designing Embodied Conversational Agents (2000).

[Churchill et al.] Churchill E.F., Cook L., Hodgson P., Prevost S., and Sullivan J.W.: "May I Help You?": Designing Embodied Conversational Agent Allies.

[Druin et al. 1999] Druin A., Montemayor J., Hendler J., McAlister B., Boltman A., Fiterman E., Plaisant A., Kruskal A., Olsen H., Revett I., Schwenn T., Sumida L., and Wagner R.: Designing PETS: A Personal Electronic Teller of Stories (1999).

[Paiva et al. 2001] Paiva A., Machado I., and Prada R.: Heroes, Villains and Magicians: Dramatis Personae in Virtual Environments. In Intelligent User Interfaces, ACM Press (2001).

[Scherer 2000] Scherer K.R.: Emotion effects on voice and speech: Paradigms and approaches to evaluation. Presentation held at ISCA Workshop on Speech and Emotion, Belfast (2000).