# Trackside DEIRA: A Dynamic Engaging Intelligent Reporter Agent

François L.A. Knoppel, Almer S. Tigelaar, Danny Oude Bos, Thijs Alofs, Zsófia Ruttkay
Human Media Interaction, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands
{f.l.a.knoppel, a.s.tigelaar, d.oudebos, t.alofs} @student.utwente.nl,
z.m.ruttkay@ewi.utwente.nl

## ABSTRACT
DEIRA is a virtual agent commenting on virtual horse races in real time. DEIRA analyses the state of the race, acts emotionally and comments about the situation in a believable and engaging way, using synthesized speech and facial expressions. In this paper we discuss the challenges, explain the computational models for the cognitive, emotional and communicative behavior, and account on implementation and feedback from users.

## Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Language Generation and Speech Synthesis; J.4 [**Computer Applications**]: Social and Behavioral Sciences; H.5.1 [**Multimedia Information Systems**]: Animations and Audio output; I.3.7 [**Three-Dimensional Graphics and Realism**]: Animation

## General Terms
Algorithms, Performance, Design, Experimentation, Human Factors, Languages.

## Keywords
Intelligent virtual agent, multimodal communication, emotion modeling, facial expressions, synthetic speech.

## 1. INTRODUCTION
Virtual Agents or Embodied Conversational Agents (ECAs) [1] resemble humans in their embodiment and cognitive and communication capabilities. Their application domain covers presenters of news or weather forecasts, sales assistants, medical consultants, educational tutors or multi-person simulation environments for trainings etc. It has been proven that people react to virtual humans basically in a similar way as they react to real ones, even if we are far from understanding the influence of different design parameters in different application scenarios [2][3]. A virtual agent should fulfill some – often conflicting – goals such as teach and/or entertain the human user [4].

Due to the relative infancy of the field, and the enormous challenges in reproducing the complexity and richness of human communicative behavior, much research has been addressing singular aspects of the multi-modal presentation, such as believable gaze behavior, faithful facial expression, lip synchronization, etc. Another group of researchers, often with a strong AI background, is working on modeling the cognitive and emotional aspects of virtual agents, endowing them with emotions, goals and problem solving. Usually, the particular applications benefit from a huge AI machinery and a lot of tailored work for the presentation skills. Because of this complexity and the different application domains it is hard to compare different ECAs. It is common practice in many sub-fields of AI, like search or natural language understanding to test one's own system on some benchmark problems, and thus compare the pros and cons of different approaches. For the community of ECA researchers, there are no similar benchmarks.

Our application, a virtual horse race reporter[1], extends the above outlined state of ECA research in the following aspects:

- The virtual reporter is an application to test reactivity and engagement. It can be triggered by generating race event sequences by a race simulator and not by still error-prone modules that try to understand user input;
- The domain of race reporting is well defined, limited but at the same time rich enough to pose challenges to both the cognitive and the presentational capabilities of a virtual reporter;
- The virtual reporter has to react to a rapidly changing dynamic environment, making time a critical factor for his assessment and presentation skills.

Because of these characteristics, the Race Reporter was posed as a challenge for the ECA community. The explicit goal of the reporter was to provide an engaging and entertaining experience by reporting on virtual horse races. In the framework of the GALA contest[2] our agent was reporting real-time on a race script supplied on the spot. Another race reporter was present, which made comparison of the reactions and overall performance possible [5].

**Cite as**: Trackside DEIRA: A Dynamic Engaging Intelligent Reporter Agent, Knoppel, F.L.A. et al., *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, Padgham, Parkes, Müller and Parsons (eds.), May, 12-16., 2008, Estoril, Portugal, pp. 112-119.


[1] In our encounters with real horse race enthusiasts, we learned that the correct term for our system is either "announcer", "race caller" or simply "commentator". Due to the proliferation of the term reporter in various documents upon learning this, we have decided to leave this denomination unchanged.

[2] See http://hmi.ewi.utwente.nl/gala/racereporter

## 2. RELATED WORK

There has been a good deal of work to provide commentary in (virtual) sports applications, both in the game industry and on an academic level. Games simulating popular televised sports usually contain some sort of audio commentary to increase user enjoyment, with soccer games being best known for it[3]. The commentary systems employed are characterized as having very little intrasentence variation and a great deal of emotion in the speech as they use output sequences of *prerecorded* sentences. Yielding more complexity and flexibility are systems designed in an academic context, such as the three RoboCup reporters described in [3][6] and the soccer reporter described in [7]. For the latter, focus was limited to facial display of emotions. The RoboCup reporters differ from our system in that those systems required a great deal of time be devoted to the event detection mechanisms, whereas we paid much attention to verbal variety of utterances and expression of tension and emotions in speech and facial expressions. Somewhat more loosely related is a project generating commentary using player-like agents in first person shooter games [8]. Here, focus was on how to add the agents without compromising gameplay.

The following sections contain our account on the design, implementation and testing of DEIRA, our virtual race reporter. Firstly, we comment on the most significant results of our domain analysis. In section 4 we then discuss the design and implementation of our system. Section 5 is devoted to evaluations of our system in terms of feedback from horse race enthusiasts as potential 'users' of the system. Finally, in section 6 and 7, we present our conclusions and outline possibilities for further work.

## 3. DOMAIN ANALYIS

As the very first step of the entire project, we investigated the domain of horse racing. This research consisted of looking at the characteristics of a typical horse race, watching recordings of actual horse races and listening to real reporters and communicating with actual horse race fans to elicit what such an audience would expect from a virtual reporter agent.

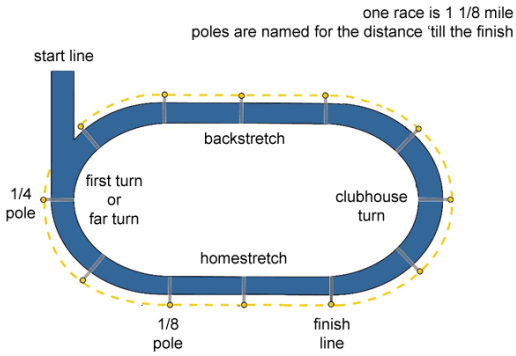



**Figure 1. Typical horse racing track**

### 3.1 Horse Race Analysis

Looking into the characteristics of horse races, it became clear that there are quite some variations in track length, track layout, number of participants etc. The variation that we found most suitable, considering the generality of the track layout and as it matches the recordings we had access to, is depicted in figure 1. The track contains a number of notable turns and stretches and is 1 and 1/8th mile long.

As we were developing our system within the context of the GALA challenge, the characteristics of the track had to be used in the context of the Horse Race Simulation Software (HRSS) provided by GALA[4]. The HRSS simulates a race on a straight track of 2000 meters with each horse racing in a separate lane. To bring in the variety of actual race tracks, elements of our typical track such as the turns and stretches were mapped onto the straight HRSS track.

### 3.2 Horse Race Commentary Analysis

The most revealing part of the domain analysis was watching, transcribing and annotating eight (US) horse races, acquired via the website of the New York Racing Association[5]. The annotation concerned assigning a value to the moments during the race when experienced tension changed, as well as writing down perceived emotional responses of the reporters.

We found that reporters tend to contract sentences as to provide more information in less time. A good deal of standard phrases with only slight variations could be discovered, which were thus fairly straightforward to capture in a grammar at a later point. Phrases such as "x is only y yards away from z" exemplify this. We also deduced a set of categories like 'overtake' and 'position report' in which the different utterances could be grouped.

The emotional content of the reports was limited, but there were some instances in which surprise, amusement, and pity were present. The excitement of the reporter, reflected in the experienced tension was the most prominent aspect in all of the analyzed reports. It builds up gradually during the race, reaching a peak when the horses come near the finish line, but also shows a characteristic peak at the start of the race. When analyzing the varying parameters of the speech, a strong positive correlation was found between perceived tension and the speed, pitch and volume of the speech. Comparing our annotations, it was clear that we all perceived higher levels of tension as the reporter spoke faster, louder and at a higher pitch.

### 3.3 User Expectations

A participation request for an online enquiry containing questions regarding the preferences of users in horse race commentary was posted on the most prominent horse race forums on the Internet. Nine actual horse race enthusiasts, who are the potential target audience for our application, supplied their expert opinion. The people who filled out the enquiry were 40 to 70 years old, generally male, with 16 or more years of interest in horse racing, and had seen races in the US, UK, and Ireland.

The most interesting results concerned the specific information needs during the race, being: the initial ordering of the horses after the start and at the finish, changes in order, noteworthy potential changes in order, total race time at checkpoints, remarkable gaps between horses, and extraordinary events.

The enquiry participants had a very distinct expectation as of the physical characteristics, the personality and the behavioral aspects

[3] See for instance: http://www.fifa08.ea.com/

[4] See http://hmi.ewi.utwente.nl/gala/racereporter

[5] See http://www.nyra.com

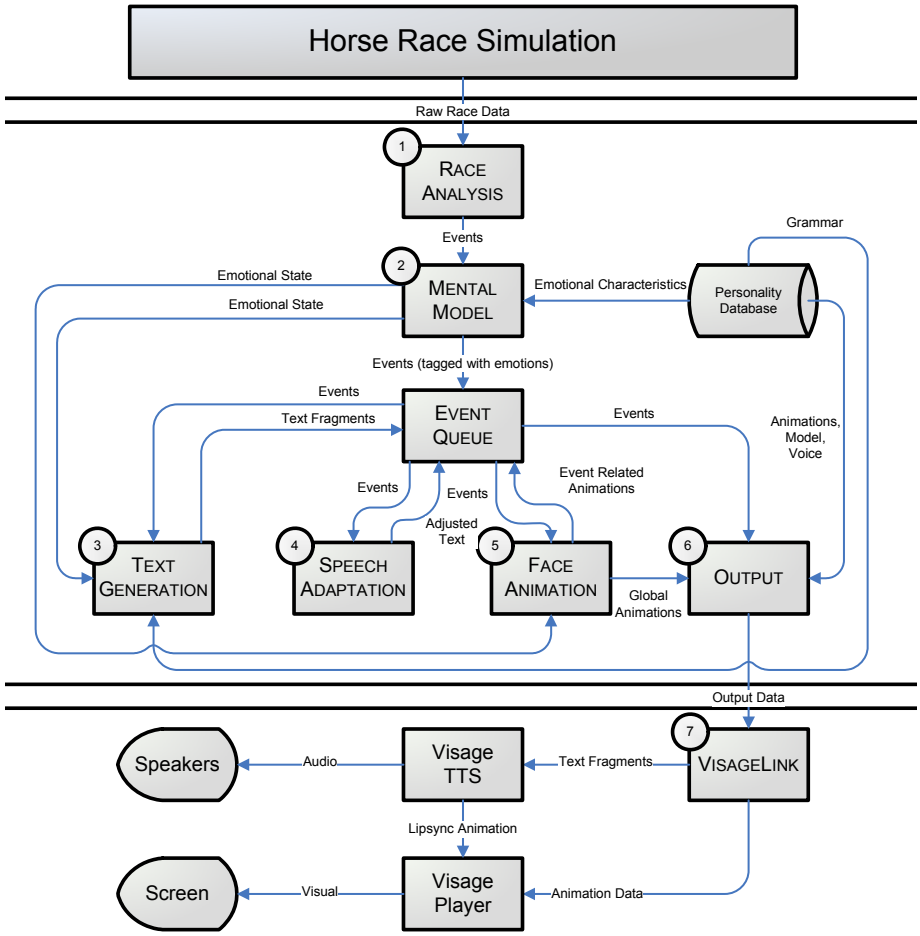



**Figure 2. System overview**

of a virtual horse race reporter. The (strong) general preference was for a masculine, middle-aged male with a regular British accent, having a neutral to authoritarian demeanor and who would not show any sign of favoritism. And lastly, not unexpected but important to note, the enquiry participants wanted a virtual horse race reporter to be as realistic as possible.

# 4. SYSTEM DESCRIPTION

With the results of our domain analysis in mind, we designed and implemented a system with a modular architecture which we will describe in the following subsection. The subsections after that contain brief descriptions of the main parts of our system. Each part of the system focuses on covering a specific aspect of the agent, such as a providing a basis for agent reasoning, generating emotions and generating corresponding facial and verbal expressions.

## 4.1 Architecture

The diagram in Figure 2 gives a global impression of the components of the system. The Horse Race Simulation Software (HRSS) and VisageLink[6] are external to the core DEIRA application. The position and speed of the horses in the race is sent from the HRSS to the Race Analysis Module (RAM) once every second (1). The RAM in turn generates events corresponding to what is happening in the race and passes these to the Mental Model Module (MMM) which determines the emotional impact of an event (2) after which it puts the event on the Event Queue (EQ). Here all other modules can get access to the available events in a synchronized fashion.

A rule set, designed by hand and based on our transcripts of actual commentary, produces an *initial importance* and an *importance decay factor* for each event the EQ maintains *importance levels* for the events, ensuring that the events with the highest current importance will be offered to the modules first. This results in *event prioritization*: the events most interesting to be reported on are processed first. Although the events are processed in a more or less linear fashion, we explicitly chose for synchronized access to a central queue module for multiple reasons. The first reason for that is to make human-like response possible in the sense that the agent could be commenting on some event, but also be reacting emotionally to something that occurs simultaneously. Secondly, modules can run concurrently, processing whatever events available in the processor time they are allotted. A final reason is to enable straightforward revoking of certain events (e.g. commentary on current rankings which have been invalidated by an overtake) in a manner that wastes as little processing time as possible. Besides the revoking, the EQ also decreases chances of repetition by lowering the importance of events very similar to a previously uttered event.

After events are put in the queue they are processed in succession by several other modules (3-5), generating text to be uttered and facial animation to display, which are finally outputted in a fused manner (6 and 7). The DEIRA system thus monitors the race

[6] For more information concerning Visage:

See http://www.visagetechnologies.com

continuously, takes raw race data as input, interprets it, and transforms this to representation suitable for controlling a human-like virtual agent. The result is a talking head uttering appropriate speech fragments and expressing emotions through facial animation and in speech characteristics.

## 4.2 Race Analysis

The RAM is responsible for detecting events occurring in the race based on the positions and speeds provided by the HRSS. Based on simple rules, in the format of *expression -> event*, new race information is provided every second. Specific events are generated for each expression that evaluates to *true*.

EXAMPLE 1. *A rule definition*

```
    ((#1.curr.pos - #2.curr.pos) < 25)
&& (#1.curr.speed < #2.curr.speed)
-> GAP; SMALL; <#1,#2>; 3.4; 2;
```

*This rule expresses that if the horses ranked first and second come within 25 meters of each other and they are getting closer, a ' small gap' event is generated.*

*Expressions* can use information about each of the horses, like position, speed, and rank and can include basic mathematical operations.

*Events* consist of a general type (GAP in Example 1), a subtype (SMALL), specific actors (the horses that are coming very close to each other), and any additional parameters (N/A for the specific event in Example 1). Two other vital parameters are always included with each generated event: importance (value 3.4 in the example) and importance decay (2 in the example). These values represent the human tendency to attach a different importance to different events depending on their relevance, and that the importance decays over time as the event becomes less interesting. The importance and decay are used both in the priority queue as well as in determining the emotional impact.

Using this rule-based approach, a large variety of events can be defined. Currently the system detects: starting of the race, speed-ups, speed-downs, stops, a horse closing in on another horse, large gaps between two subsequent horses, overtakes, the passing of certain positions of interest on the track (e.g. quarter pole), and the finish of the race. If no events are detected for some time, a background event is generated to trigger the reporter to fill in the silence with general information on the participants.

## 4.3 Personality, Emotion and Emotional State

The characteristics concerning the body and mind of the reporter are defined in the Personality Database (PDB). The reporter's personality is expressed in terms of a specific generative grammar, a voice, a face and emotional characteristics. Each of these can be altered or replaced to give the reporter a different personality.

There are four distinct emotions we have considered contributing to the emotional state of the reporter, which are excitement (or tension), surprise, amusement and pity. The abovementioned *emotional characteristics* in the personality database specify how strongly these aspects of the emotional state of the agent are influenced by events that are processed. The influence the characteristics in the PDB have on these aspects is basically a numerical multiplication of the emotional impact assigned to a certain event. For each emotional aspect, a different factor is present. All together, these factors represent the tendency for certain emotions of the reporter.

The dynamically changing, actual emotional state of the reporter is reflected in intensity values for the defined emotional aspects. The MMM updates these values by accumulating the emotional impact of all events that were generated during the race so far, taking decay factors, that are employed to mimic the fading of emotions, into account. The emotional impact of the events as well as the decay of that impact is determined based on the importance characteristics and the type of the event, as specified by the RAM. A very specific influence on the emotional state of the reporter is the remaining distance to the finish, which is inversely related to the base excitement level of the reporter.

The emotional model of the reporter has been designed to accommodate for the display of emotions present in the actual commentary we analyzed. With the excitement level being the main factor in that emotional display, we focused on having the system correctly display excitement. Very notable emotional state changes that we found are, for instance: the start of the race inducing a briefly raised excitement level and the end of a race inducing a drop in excitement level. Despite having the focus on excitement level, we have designed the MMM in a way that allows the existing emotional parameters to be tuned such that they have a more pronounced influence on the output as well as being able to add more of these parameters.

## 4.4 Verbal Expression

The generation of speech output is one of the key functionalities of the system. This process is divided over several components within the system, being the Text Generation Module (TGM), the Speech Adaptation Module (SAM), the Output Module (OM) and the external component VisageLink. In this process, the TGM is responsible for determining what the agent could say about an event and deciding on what he will say about an event if time allows. Upon receiving an event, the TGM generates a set of potential output sentences based on the RAM event description.

These sentences are constructed using a generative context-free grammar supporting variables and conditionals. That grammar consists of a set of production rules that are elaborated in a recursive manner using the event description as input for the elaboration. All rules applicable to that event are exhaustively transformed into sentences. All of the generated sentences have identical semantics but vary in syntax, thus providing the agent with a range of textual options to supply the event information to the user.

To enable generation of sentences that are applicable for specific emotional states the production rules can be augmented with boundaries on the values of the emotional state for which they are valid. Currently, only bounds on the tension levels are supported.

An example input event could be 'GAP; SMALL; <Participant 1, Participant 2>', with the participants being objects. Using information from those objects and information on what variables occur in the sentences for this specific event, variables are instantiated yielding for instance {ACTOR2} = "Azure" and {YARDS} = 7.

EXAMPLE 2. *Elaborated rule set*

| | |
|---|---|
| GAP | \|= -GAPTYPE = SMALL-, [Dec-SmallGap] |
| Dec-SmallGap | \|= {ACTOR2}, [DG-Action], {ACTOR1} |
| DG-Action | \|= 'very tight behind' |
| DG-Action | \|= 'coming within', [Gapsize-Plural],'of' |

*These rules produce sentences like: 'Azure very tight behind Eben', 'Azure coming within seven yards of Eben' and a number of similar sentences..*

When all possible sentences for a certain event have been generated they are assigned a score reflecting: the applicability to the emotional state at that time, occurrence in text-selection history and a random influence. The mentioned applicability is the output of a distance function reflecting the distance between the current emotional state of the reporter and the intervals on the emotional dimensions for which a sentence is deemed valid. For instance: if a sentence is valid for tension levels between 3 and 5 and the current tension level of the reporter is 1, then the score of that sentence will be negatively influenced. The text-selection history keeps track of all the phrases that occurred in sentences that have been selected *by the TGM* for uttering. These phrases are then assigned penalty scores which diminish over time. This approach results in the agent disfavoring sentences containing phrases that have been previously used, leading to ample variation in the textual content.

The SAM transforms the generated text in such a manner that the voice chosen for output pronounces the different words correctly. It is also responsible for determining at what speed, pitch and volume the agent should utter certain sentences. As our initial analysis of actual horse race commentary showed: the three mentioned variables are directly linked to the excitement experienced. This is reflected in the values we assign for those variables every time a sentence is uttered. The final part of speech generation is feeding the input data to a Text-To-Speech engine. This is done by the OM, which will be discussed in 4.6.

## 4.5 Non-verbal Expression

In the final step, the non-verbal signals are determined which are to be displayed by the head model. This is done by the Facial Animation Module (FAM) that translates the emotional state of the reporter, as expressed by the Mental Model Module, to dynamic facial expressions. The animations themselves are stored as pre-designed MPEG-4 FBA files [9] and can be applied to any head model. The FAM animates the face with two separate layers of animation.

The *first layer* consists of basic `idle' head movements. At fixed intervals a new basic head movement animation is selected based on the average excitement of the reporter (combined with certain heuristics) since the last interval. A higher excitement leads to more pronounced movements of (parts of) the head. Effects that become more visible include saccadic eye movements, small motions of the head, jaw grinding, expansion of the nostrils, and contraction of other facial muscles.

The *second animation layer* is constructed from one or more head animations related to the emotional state of the reporter as it is at the moment of an occurrence of an event. The animations involved in this layer are more pronounced than those in the first. The animations include expressions such as smiling and frowning, but also occasional eye blinking and looking in a certain direction, with all animations being more pronounced as excitement rises. At special moments in the race, such as the start and finish, the head turns to look directly into the `camera', which is also controlled by this layer.

The first layer animations are sent directly to the visualization front-end because these idle animations are not dependent on the events that occur, but are simply there all the time. Second layer animations are added to each event and become visible only if the event reaches the output stage. Note that lip synchronization is not taken care of by the FAM since this is done by the Text-To-Speech (TTS) engine and Visage (see Visage TTS in figure 2, after step 7).


**Figure 3. System screenshot**

## 4.6 Output

The previous paragraphs described the step-by-step process in which events are tagged with various pieces of information concerning emotions, textual formulation and non-verbal display.

The last step is fusing all this information together and producing synchronized speech and animation. This is the task of the OM and the external VisageLink application. Information is passed between these applications using a socket. The OM first retrieves the head and voice to use from the PDB and sends this to the VisageLink application. Then for each event the OM transfers the text to be uttered including speed and volume characteristics as well as any event specific animations to be shown. The VisageLink application then animates the head and transforms the text into speech using a TTS engine.

We used Nuance's RealSpeak(TM) US English voice[7] for this purpose which is Microsoft® SAPI 5.1 compatible. The head model was edited using tools that are part of the Visage SDK and transformed into the Visage AFM format [10].

We ran performance tests on our system and found that the minimal delay between the creation of an event to it actually being output to VisageLink is about 30 milliseconds. Of course in many cases events are held back until the reporter finishes reporting on a previous one. Events *that are actually sent for output* sit in the queue for about one second on average and never longer than five seconds.

Figure 3 shows the entire system in action. At the top the VisageLink application shows the reporter. In the middle the HRSS is shown and below that, the Trackside DEIRA application displays both the generated sentences and emotional state.

# 5. USER STUDIES

During the entire iterative development and implementation process, evaluating the system with potential users provided the feedback for subsequent improvement. The horse race enthusiasts we came into contact with for the initial domain analysis served as our most important group of test users in these evaluations. Although the implementation differences between the iterations of our system are quite notable, discussing the most important findings of the intermediate evaluations is relevant for the system we've introduced in detail. It is furthermore interesting to see how the changes we made to the system over time have influenced user experience. After a short introduction of the evaluation setup and summary of the most interesting intermediate results, we will discuss the evaluation of the current system.

## 5.1 Prior Evaluations

The evaluations carried out in the earlier stages focused on a number of aspects of our reporter, such as the visual characteristics, audio characteristics, content and timing of the utterances etc. Users were presented with an online questionnaire containing three sections of open and Likert-scale questions, interleaved with instructions to view movies of the system in action. All movies in the evaluation have been generated using the same race script as input to the HRSS.

Important to note is that the movies were designed with the GALA requirements in mind: Of the two movies, the first movie contained *only visuals of the virtual race* combined with audio commentary and the second movie contained *only visuals of the reporter* and audio commentary. The first movie was thus meant as a more common representation of horse race commentary and the second one was aimed at eliciting a positive user experience solely from looking at and listening to the reporter.

The most elaborate of the intermediate evaluations done in this manner concerned the system as it was running using XFace[8] instead of Visage as output application and utilizing our own database of pre-recorded speech fragments instead of a TTS engine. The details of the differences are not interesting at this time, but the implications of these differences are important to keep in mind as we compare results between evaluations.

In total, five horse race enthusiasts and ten users with little to no experience with horse races participated in this evaluation. In general, the answers given by horse race enthusiasts and the inexperienced users did not differ notably.

For sake of brevity, we can only discuss a few results. The complete discussion of this evaluation is available via the DEIRA website[9]. One of the notable results is the extent to which test subjects found our reporter to mimic a real reporter. This was about 33% and furthermore, all users stated that they greatly preferred a human reporter over our reporter. The audio content was deemed understandable and believable by over 80% of the users, but variety in both content and intonation was not up to standards. More than half the test users found that the reporter did not display emotions often and intensely enough.

The users' feedback led to the following recommendations for major improvement of the system:

- Improve timing of commentary
- Add more background information to the commentary
- Add more variety in the content of speech
- Add more variety of intonation
- Change voice to a more British sounding accent
- Improve lip synchronization
- Add more movement in the face
- Increase visual emotional expression

The first six improvements were important reasons to switch to a TTS-based system. Doing so would greatly decrease the labor required to extend the vocabulary of the system and increase the manipulability of the speech output. Although properly pre-recorded speech samples sound more natural, varying content, pitch and speed requires a lot of different samples.

The positive effects the transition to a TTS-based system would have, combined with a diminished minimal delay between event generation and the report on that event led us to adapt the system for use with Visage and a TTS engine. This adapted system is the current version of DEIRA. After our final evaluation we were able to determine whether the implemented changes have truly influenced the user experience in a positive way, as we will discuss next.

[7] The UK English version did not output viseme information, necessary for lip animation, when the choice for a specific TTS engine was made. See also http://www.nuance.com/realspeak for more information.

[8] See http://xface.itc.it for more information

[9] See http://www.queequee.net/deira

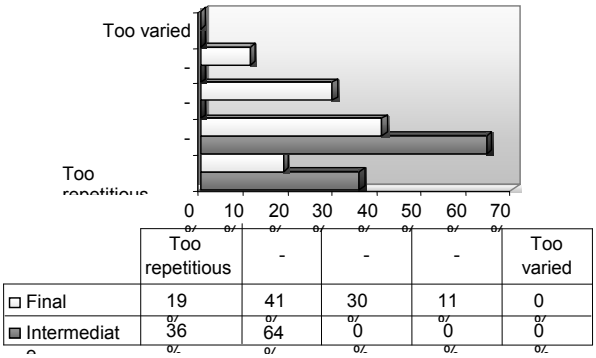
## 5.2 Final Evaluation

The last evaluation of our system was conducted in exactly the same manner as the evaluation before it, thus enabling straightforward comparison of the quantifiable answers.

An important difference between this evaluation and the abovementioned one is in the level of horse race enthusiast participation. We received 25 answer series from inexperienced users and only two from horse race enthusiasts in this evaluation. Worth noting is that the two responses of horse race enthusiasts we did get were relatively negative. Considering the lack of such a difference in earlier evaluations we believe this is also partly due to the overtaxing of the online horse race enthusiast community. Be that as it may, a number of differences and notable similarities between the results of the evaluations are clearly visible.

Relatively more users rated the audio commentary as not (so) understandable. Whereas in the previous evaluation 93% of the users rated the commentary as intermediately to very understandable, this time the number has gone down to only 74%.

Judgment of variation in content improved, as is visible in the graph below. This – actually, even a more pronounced improvement – was expected.



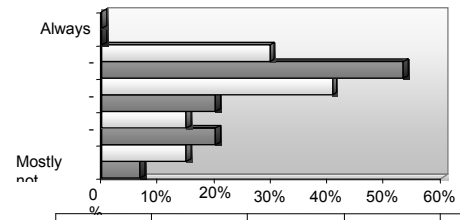

| | Too repetitious | - | - | - | Too varied |
|---|---|---|---|---|---|
| Final | 19 | 41 | 30 | 11 | 0 |
| Intermediat | 36 | 64 | 0 | 0 | 0 |

**Figure 4. Variation in speech content**

When looking at the results concerning variation in intonation, frequency of emotion in speech, and the intensity of emotion in speech, no real differences are visible. The general consensus is that both pre-recorded speech samples as well as our employed TTS engine fell short in these aspects.

Interestingly, only two users found that the reporter had not reported all he should have reported. On the other hand, six users found certain sentences not to be "logically structured". On closer inspection, almost all 'examples' given by users contained only comments on either the speed and resulting lack of comprehensibility of the audio or the coherence of the discourse.

We expected that, once the timing of reports had greatly improved, so would the user experience with regard to timing issues. The responses to the question "Were the comments on events expressed at the right moment?" showed that contrary to the expectation, users were less satisfied with timing of comments. The graph below illustrates this difference.



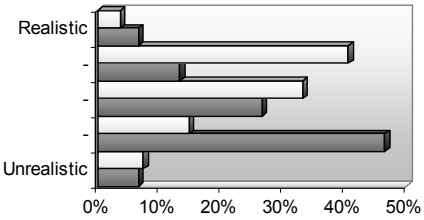

| | Mostly not | - | - | - | Always |
|---|---|---|---|---|---|
| Fina | 15% | 15% | 41% | 30% | 0 |
| Intermediate | 7 | 20% | 20% | 53% | 0 |

**Figure 5. Correctness of timing of reports**

It is hard to interpret this result, but a possible explanation is that as in the final version more events were reported on, event timing correctness was perceived more critically. Especially reporting on background information at inappropriate times could be cause for the negative reactions, as one of the users commented: "they dont comment on jockeys previous experiences in mid race!"

With regard to whether users would prefer our reporter over a human one or no reporter at all, no difference between the results of the evaluations is present. People greatly prefer a human reporter but do favor a virtual reporter to none at all.

A very striking result is the users' perception of the realism in the visual appearance of the reporter as the following graph shows.




| | Unrealistic | - | - | - | Realistic |
|---|---|---|---|---|---|
| Final | 7% | 15% | 33% | 41% | 4% |
| Intermediate | 7% | 47% | 27% | 13% | 7% |

**Figure 6. Realism in visual appearance**

The difference is clearly visible in the graph. There are a number of possible explanations for this difference, one of them being that some participants may have seen the reporter with a 'squashed' face due to an issue with the aspect ratio of the movie. The result could also be attributed to the improvements we have made, including new animations, textures and stance of the head, but due to the abovementioned technical issue we cannot conclusively do this. There were no noticeable differences between the visual rendering quality of XFace and Visage, which excludes such differences as a possible cause of the different values.

With regard to the other visual aspects, no significant differences between both evaluations can be found. What *is* very interesting though is that the average score on the question "Grade the reporter in resembling a real reporter" differs quite a great deal between evaluations. In the final evaluation, the score has risen from 33% to 49%, signifying a clear improvement in the system. It is important to note that for both scores very little difference was present between the scores given by the horse race enthusiasts and the inexperienced users.

## 6. CONCLUSION AND DISCUSSION

We presented DEIRA, a highly reactive virtual horse race reporter, capable of interpreting raw race observation data, conveying his emotional state and commenting on race events by speech accompanied by facial expressions in real-time. The system uses rule-based technology for event analysis and mental models, natural language processing and a third-party TTS engine and facial model for output synthesis.

DEIRA can serve as a basis for other similar applications where real-time reactions and engagement are of major importance. The modular design of the system and the separate declaration of the domain-specific knowledge make it easy to adapt it to different domains or user groups (e.g. language, age group). It would be very straightforward to create a virtual reporter for other types of linear races like virtual car racing. Additionally, our system could be adapted with ease to create a virtual commentator for domains such as the popular first person shooter games. To achieve this, the engine to process events could be used as is. The only challenge would lie in generating proper events and parameters and defining the appropriate event-to-text grammar.

We have received positive feedback from professional circles as of the mental and communicational capabilities of DEIRA and the overall engaging experience. Evaluation results show that potential users are impressed by the capabilities of the system. They preferred having our virtual reporter commenting on a virtual race, as opposed to over a race system without one, signifying a clear added value of our system. The same people were still critical in comparing the virtual reporter to a real one.

Contrary to our expectations, users did not value improvements in the timing of comments and facial appearance of the reporter. Looking at the system as a whole, we can think of two explanations for this. First of all, the synthesized speech used in the final version may have overcasted other, improved aspects of the reporter. Similar cross-modality effects have been previously reported on in connection with virtual agents [4].

Secondly, our two-stage evaluation results could also have to do with the 'uncanny valley' effect. Mori originally stated that, for robots, realism works against acceptance and user satisfaction: people become more critical towards artificial creatures which look like real humans, as this increases the expectations of the users. This effect has been a major concern for the virtual agents community [11].

Keeping these options in mind, one can think of further work along two lines. One option is to improve subtle aspects of the facial model (e.g. add hair) and behavior to get a better illusion of realism. Improving the quality of the idle motions and improving gaze behavior would be steps in this direction. Also, more articulated expression of emotions could increase engagement and user satisfaction. Last but not least, improved quality of synthesized speech is crucial. With SAPI 5.3 TTS voices that support SSML becoming available the SAM could be greatly extended to include variation in emphasis, intonation and thus increasing expressivity in the speech.

In line with the uncanny valley effect is the option to choose for employing a less realistic head model, but with the possibility of enhanced expressivity, including effects known from the traditional animation films. Exaggerated facial expressions like bulging eyes, together with effects like hair raising could potentially increase the entertaining and engaging quality of the virtual reporter. However, with such a choice we would deviate from the initial challenge of reproducing the behavior of a real reporter.

## 7. ACKNOWLEDGEMENTS

We would like to thank the GALA organization, all those that were willing to participate in our enquiry and our friends and family. We furthermore want to thank Prof. Igor Pandzic for his cooperation in our use of Visage.

## 8. REFERENCES

[1] Cassell, J. 2000. Embodied conversational agents. Cambridge, Mass: MIT Press.

[2] Nass, C., Steuer, J., and Tauber, E. R. 1994. Computers are social actors. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating interdependence (Boston, Massachusetts, United States, April 24 - 28, 1994). B. Adelson, S. Dumais, and J. Olson, Eds. CHI '94. ACM Press, New York, NY, 72-78.

[3] Andre, E., Binsted, K., Tanaka-Ishii, K., Luke, S., Herzog, G. and Rist, T. 2000. Three RoboCup simulation league commentary systems. AI Magazine 21 (1), 57-66.

[4] Ruttkay, Zs., Pelachaud, C. (Eds.), 2004. From Brows to Trust: Evaluating Embodied Conversational Agents, Kluwer.

[5] Strauss, M., Kipp, M. 2008. ERIC: A Generic Rule-based Framework for an Affective Embodied Commentary Agent. Accepted at AAMAS08.

[6] Andre, E. and Rist, T. 2001. Presenting through performing: on the use of multiple lifelike characters in knowledge-based presentation systems. In Knowledge-Based Systems, Volume 14, Issues 1-2, 3-13.

[7] Bui, T.D., Heylen, D.K.J. and Nijholt, A. 2004. Building embodied agents that experience and express emotions: A football supporter as an example. In Proceedings of the 17th annual conference on Computer Animation and Social Agents (CASA2004), Geneva.

[8] Fielding, D., Fraser, M., Logan, B., and Benford, S. 2004. Extending game participation with embodied reporting agents. In *Proceedings of the 2004 ACM SIGCHI international Conference on Advances in Computer Entertainment Technology* (Singapore, June 03 - 05, 2005). ACE '04, vol. 74. ACM, New York, NY, 100-108.

[9] ISO/IEC 14496-2. 2001. Information Technology -- Coding of audio-visual objects -- Part 2: Visual.

[10] Pandzic, I. S. et al. 2003. Faces Everywhere: Towards Ubiquitous Production and Delivery of Face Animation. In Proceedings of MUM03, Norköping, Sweden.

[11] MacDorman. K.F. 2005. Androids as an experimental apparatus: Why is there an uncanny valley and can we exploit it? CogSci-2005 Workshop: Toward Social Mechanisms of Android Science, 106-118.