VHML – Virtual Human Markup Language

Andrew Marriott

School of Computing, Curtin University of Technology,
Hayman Rd. Bentley, Western Australia.
raytrace@cs.curtin.edu.au
http://www.vhml.org/

Abstract. This paper details the research being done at Curtin University in creating a Virtual Human Markup Language (VHML) that allows interactive Talking Heads to be directed by text marked up in XML. This direction makes the interaction more effective. The language is designed to accommodate the various aspects of Human-Computer Interaction with regards to Facial Animation, Body Animation, Dialogue Manager interaction, Text to Speech production, Emotional Representation plus Hyper and Multi Media information. This paper also points to audio and visual examples of the use of an early version of the language. VHML is currently being used in several Talking Head applications as well as a Mentoring System. Finally we discuss the future development of VHML. The VHML development and implementation is part of a three-year European Union Fifth Framework project called InterFace.

1 Introduction

The Virtual Human Markup Language (VHML) uses / builds on existing (de facto) standards such as those specified by the W3C Voice Browser Activity, and adds new tags to accommodate functionality that is not catered for. The language is XML/XSL based and consists of the following sub-systems:

- DMML Dialogue Manager Markup Language
- FAML Facial Animation Markup Language
- BAML Body Animation Markup Language
- SML Speech Markup Language
- EML Emotion Markup Language
- GML Gesture Markup Language

The intent of VHML is to facilitate the realistic and natural interaction of a Talking Head/Talking Human (TH) with a user such as can be found in Pandzic (2001). Previously developed TH systems have been seen to be acceptable as an HCI (Beard et al., 1999) with effective use of the system being detailed in Marriott (2001c) and an improved version using an early version of VHML in Marriott et al (2001b).

2 The Problem

In figure 1, the information returned from the server Knowledge Base should be marked up in such a way that it is easily parsed, searched, categorized, etc and this implies a consistent markup language. Also the final rendering and delivery of the marked up text depends upon whether the output scene is text only, audio only, face or full body and could be in many forms:

- a straight textual interface for a very low bandwidth interface,
- an interactive Web based multi and hyper media display,
- a voice enabled system which lets the user hear the answer,
- a TH with complex facial gestures and voice which shows the personality and emotion of the TH,
- an entire synthetic human figure with body language.

The information markup should stay the same but the way in which it is displayed should change dependent upon the form of output - textual, vocal, Talking Head, Body Language.

It should also be possible to facilitate the direction of a Virtual Human interacting with a user via a Web page or stand alone application. For example, a Virtual Human that has to give some bad news to the user - "I'm sorry Dave, I can't find that file you want." – may speak in a sad way, with a sorry face and with a bowed body stance. In a similar way, a different message may be delivered with a happy voice, a smiley face and with a lively body. XML and XSL provide this technology. VHML tags such as `<smile>, <anger>, <surprised>` have been specified to produce the required vocal, facial and emotional actions.

Finally it is necessary for the Virtual Human to behave and react in a believable and realistic manner for it to be effective and acceptable as an HCI.

3 The Solution – the Extensible Markup Language (XML)

XML is the Extensible Markup Language (W3C, 1997). It is a simplified dialect of the Standard Generalized Markup Language (SGML) that is relatively easy to learn, use and implement, and at the same time retains much

of the power of SGML. XML allows a user to specify the tag set and grammar of their own custom markup language. It has a similar format to HTML but is more strict. VHML has been designed using XML to enable the directing of Virtual Humans

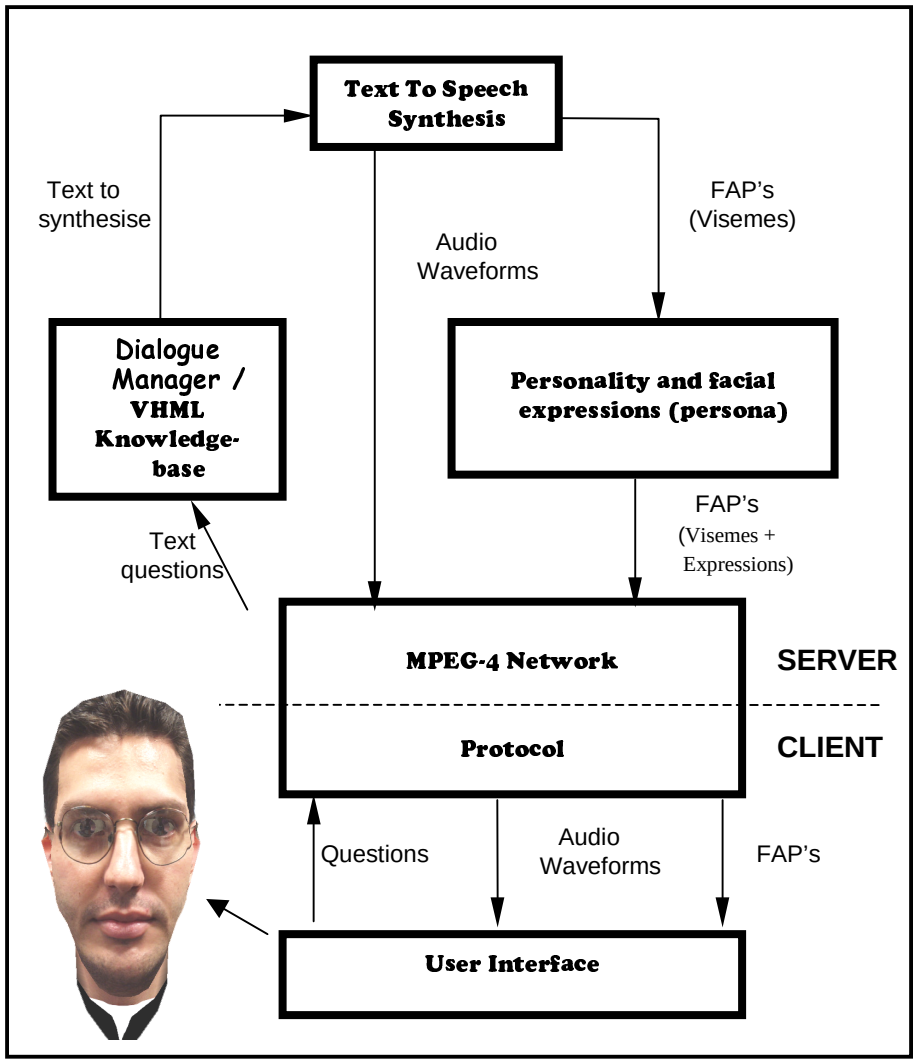

Figure 1: The flow of questions/responses inside the Talking Head User Interface

.

Detail of the design and implementation of the VHML rendering onto a Talking Head can be found in Marriott (2001b). A preliminary document specifying the structure of VHML can be found in VHML (2001).

Figure 2 shows the Knowledge Base response to the user enquiry "What are you?" and has been marked up with a VHML sub-language – DMML. It would become plain text after the substitution of the XML atomic tags.

```
<first_name/>, <welcome/>. I am <mentorName/>.
I was developed by <mentorMaster/>.
<mentorDescription/>
<mentorPurpose/>
You can find out more about me from <mentorHomeURL/>.
```

Figure 2: Segment of XML marked up text showing the use of atomic tags

In a normal interactive session, the Talking Head client will know the name of the user. This name may be used to set a Dialogue Manager variable that can be used in conversation. In this example, `<first_name/>` is the stored name of the user who is making the enquiry, `<welcome/>` is a language dependant greeting that has been set dependant upon the user's home country or domain name, `<mentorName/>` is the name of the Dialogue Manager, etc. So the final plain text may become something like:

Freda, Guten Tag. I am Mentor. I was developed by Andrew Marriott. ….

Some of the DMML atomic tags are set dynamically for each user, some are set dynamically once per user session and some are semi-hard coded into the system.

The text in figure 3 contains VHML markup tags – SML, EML, FAML and GML - that could be used to add emotional, facial and speech effects to the response. These tags and their effect would need to be suppressed for a text only display but would add emotion to a voice and/or facial rendering.

```
<sad>
 You <emph>said</emph> to me once <pause length="short"/>
 that pathos left you unmoved, but that beauty, <blink/>
 <emph affect="b" level="moderate">mere</emph> beauty,
 could fill your eyes with tears.
</sad>
```

Figure 3: Text marked up with complex XML tags to specify emotion

Audio examples of VHML marked up text can be found in JRPIT (2001a) and movies of a Talking Head using VHML in JRPIT (2001c).

Notice that the <mentorHomeURL/> in figure 2 would probably become a URL and hence the pure text may contain "http://www….". Therefore, in the "rendering" :

1. A web based display may turn that into a link
2. A vocal display may have to change the text into something like "You can find out more about me *from the link www.blah*".
3. A Talking Head may say the above and also open up a new browser window.
4. A Virtual Human may point to the link, etc.

XSL is a style sheet language designed to be used with XML data and documents to help in the different "renderings". Unlike HTML, which defines the rendering or display behaviour for each of its elements, XML says absolutely nothing about how the data is to be displayed. XSL allows the author to apply formatting operations to XML elements. XSL is a language in which the author can indicate that the <emph> element should be ignored or should be rendered (by the Face and/or the TTS).

4 Current Applications

In early stages of use of VHML are the

- MetaFace system (http://www.metaface.computing.edu.au) and the
- ***Mentor System*** (http://www.mentor.computing.edu.au/).

Both are Dialogue Managers that respond to user requests and hence there is a need for a consistent domain knowledge base so that it may be shared between systems.

With the ***Mentor System***, users may query the system in English about various aspects of their study. It will greet the user in a non-deterministic fashion, respond to various requests with varying phrasing such as "what is the weather", "what is the time", etc. Occasionally, in response to a weather update request, it will data mine a weather Web site to report accurate meteorological information, etc. A recent experiment indicated that the system correctly recognised over 75% of user requests. The system is not merely reactive but also offers timely unsolicited advice about assignments, study deadlines and time-tabling issues. Core to this is the central ***Mentor*** Dialogue Manager and Knowledge Base Manager - a Java based mini operating system in its own right

The rendering of the response can be via plain text, a Web page or Talking Head and hence the Knowledge Base is being marked up in VHML with an XSL-like transform engine as the last stage of the output processing before it is sent to the "display".

In order to further test and improve VHML, the Facial Animation research group at Curtin has completed a Talking Head Detective story - this demonstrates the emotional capability of the Talking Head and the power of VHML. The story is interactive in that the listener can solve the mystery through a question-answer mechanism.

The project will be concluded with an evaluation by the users of the Talking Head via a questionnaire that will be used to test the effectiveness of VHML in making a "better" Talking Head interface.

As part of the preparation for the story, the group has performed an inspection / verification of the implemented VHML tags, developed new tags and formulated a long-term strategy for VHML evolution as an international standard. A format (in XML) plus a Dialogue Management Tool (DMT) has also been produced for the production of dialogues (Gustavsson, 2001). That is, some stimulus from the user, typically a question, produces some output from the Dialogue Manager with some weighting or confidence level (Figure 4). The Dialogue Manager may also move into a different state having provided that response.

User : Can you tell me about Talking Heads? [STIMULUS]
DM : Yes, <*smile*>what would you like to know?</*smile*> [RESPONSE]
(DM now moves into state concerned with knowledge about Talking Heads)

Figure 4: User Dialogue plus State Change

5 Future Research

Further work needs to be done in creating more tags and fitting these to the appropriate TH persona- a phony smile will always be seen as a bad thing in a face regardless of whether the face is real or not. This low-level development of the TH tags must be accompanied by research into high-level tag classes that embody an overall feel to a body of text. Currently, the marking up of text is manually intensive, tedious and prone to error. Current research is investigating the automatic marking up of text for a given subject / context / type of presenter.

A long-term goal is for THs that can remember how certain text / words were marked up in the past and use this knowledge to transparently process minimally tagged text in a similar manner – ie. the THs learn from experience.

One deliverable of the project is a Web site that will disseminate the results of the project via a full-body Virtual Presenter who will give an interactive overview on many topics such as advances in MPEG-4 / MPEG-7 standardisation, international research on face and body animation, and speech synthesis. Therefore, it is necessary that the knowledge base is marked up with tags that reflect a full-body presentation.

6 Conclusion

The use of VHML may mean that a Virtual Lecturer or Virtual Distance Education Tutor with appropriate persona and tags is seen as being erudite and approachable, a Virtual SalesPerson in a Web page is seen as trustworthy and helpful, etc. Scenarios that use full VHML functionality will hopefully be seen as even more humane, more believable.

The TH project forms a small part of a European Union 5th Framework Project whose objective is to define new models and implement advanced tools for audio-video analysis, synthesis and representation. The human machine interface will have a face and voice, a body and gestures and will use all its artificial senses to "understand" high level messages coming from the user and will use all its virtual actuators to provide back acoustic-visual responses.

In conclusion, what is innovative and important is the possibility of directing the way in which a Virtual Human interacts with users so as to appear to be more human, humane and believable.

References

Beard, S., Crossman, B., Cechner, P. and Marriott, A. (1999), "FAQBot".In *Pan Sydney Area Workshop on Visual Information Processing*,University of Sydney, Sydney.

Gustavsson, C., Strindlund, L. and Wiknertz, E. (2001), "Dialogue Management Tool".In *Talking Heads Technology Workshop*,, Perth, Western Australia.

JRPIT (2001a), "Audio examples for "The Face of the Future". Online at http://www.interface.computing.edu.au/papers/jrpit-hci/audio".In , .

JRPIT (2001c), "Video examples for "The Face of the Future". Online at http://www.interface.computing.edu.au/papers/jrpit-hci/video/".In , .

Marriott, A. (2001c) In *Java in the Computer Science Curriculum (to be published)*(Ed, Greening, T.) LNCS, Springer, .

Marriott, A., Beard, S., Haddad, H., Pockaj, R., Stallo, J., Huynh, Q. and Tschirren, B. 2001b, 'The Face of the Future' *Journal of Research and Practice in Information Technology*, **32,** 231-245.

Pandzic, I. S. 2001, 'Life on the Web' *Journal of Software Focus*, **To be published.**

VHML (2001), "Virtual Human Markup Language. Online at http://www.vhml.org/".In , .

W3C (1997), "Extensible Markup Language (XML) 1.0. Accessed April 1997. Online at http://www.w3.org/XML/".In W3C, .