V H M L – Uncertainties and Problems. A discussion…

Andrew Marriott[1], John Stallo[2]

[1] School of Computing, Curtin University of Technology,
Hayman Rd. Bentley, Western Australia.
raytrace@cs.curtin.edu.au

[2] Microsoft Corporation, One Microsoft Way
Redmond, WA 98052-6399, USA.
johnstallo@msn.com

Abstract

This paper discusses emerging problems of VHML – the Virtual Human Markup Language. VHML is XML based and allows interactive Virtual Humans to be directed so that human-virtual human interaction is more effective. The language addresses issues related to Facial and Body Animation, Dialogue Management, Text to Speech production, Emotional Representation plus Hyper / Multi Media information. Variants of VHML are being used and evaluated in several Talking Head applications as well as a Mentor System.
VHML is an evolving language that aims for standardisation. To meet the rigorous criteria for this, VHML must be open and frank in its (self) evaluation, it must meets the needs of its targeted audience, it must be easy to learn, use and extend, etc. Peer discussions about VHML as well as a first implementation of the language from the specification have revealed problems, inconsistencies and deficiencies that need to be addressed before VHML can be seen as a stable markup language for wide spread use. This paper details some of these problems so as to promote open discussion. It also indicates the current state of VHML at specification level 0.4.
The VHML development and implementation is part of a 3 year EU 5th Framework project called InterFace.

Introduction

VHML (http://www.vhml.org/) 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 currently 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
- XHTML Hypertext Markup Language

Although general in nature, the intent of VHML is to facilitate the realistic and natural interaction of a Talking Head/Virtual Human (TH/VH) with a user (Figure 1).

```
<?xml version="1.0"?>
<!DOCTYPE vhml SYSTEM "http://www.vhml.org/DTD/vhml.dtd">
<vhml>
   <person disposition=”angry”>
      <p>
         First I speak with an angry voice and look very angry,
         <surprised intensity=”50”>
            but suddenly I change to look more surprised.
         </surprised>
         <happy wait=”2s”>
             Then I change to become very happy instead, which you both will
             see on my face and hear in my voice. The happiness was expressed
             in two seconds before I started to talk.
         </happy>
         <default-emotion>
            The happiness doesn’t last for long and now I’m angry again.
         </default-emotion>
      </p>
   </person>
</vhml>
```

Figure 1 An example of text marked up with VHML

VHML and its evaluation has been detailed elsewhere [1] [2] [3] [4] [5]. Examples of emotional audio and video of THs using an early version of VHML can be found in [6] [7].

Uncertainties

Uncertainties for users and implementers are deterrents to the use of a language. They must be clarified.

1. Semantics

   What are the semantics of the tags and their attributes? [4] has pointed out several needed definitions for the VHML specification (e.g. what exactly does the "wait" attribute mean in terms of implementation?) and these, in general, will modify the next VHML specification.

   However, a more fundamental issue is the question of what VHML actually specifies – is it just syntax or is there unambiguous semantics behind a tag? What forms the actions of <angry> or <disagree>? This issue is seen by many developers – what is the meaning of a specific viseme or expression FAP in MPEG-4 facial animation, what is the rendering of a web page given some HTML content, etc? What controls the final rendering seen by the user? There are strong arguments by both those who provide the content and want control over its rendering and by those who want the user to have control over the rendering because of some social, cultural or environmental requirements.

   The solution seems to be via "levels of implementation" – the semantics of a tag or attribute is given by **intent** as a default. For example, the **intent** of the <disagree> tag is to cause the Virtual Human to disagree in some implementation dependant way. This may be a sedate shake of the head (culturally specific), a movement of the entire body, a vigorous exaggerated movement for a cartoon-like character, etc. The **intent** is to disagree; the manner in which it is done depends upon the implementation, the type of character, the character's personality, etc. It is important to recognise that a Virtual Human should have a personality and that this should have a greater control over the final actions than some "hardwired" definition of what a tag means. So the rendering is in the hands of the implementer or the designer of the character's personality.

   However, it is also very useful for content developers to be able to exactly specify what a tag does especially in a tightly controlled environment. It is also needed if we are to allow personalities to be able to define what they want to do when they "disagree" for example.

   So a mechanism must exist within VHML that can specify, at a lower level, exactly what an action means. This introduces two new problems: how do we specify these actions in a standard way and how do we specify non-visual actions?

   To specify a low level action such as <disagree>, we must either use an existing standard or create it. Fortunately, an international standard for specifying these actions does exist – MPEG-4 Facial and Body Animation Parameters [8]. This would unfortunately tie VHML to MPEG-4 but would only be necessary (for this lower level specification) if exact content rendering were required.

   Secondly, we must be able to specify non-visual rendering exactly: what does it mean to speak in an angry voice? It is not enough to give experimental results [9], we need a specification that is unambiguous and that is potentially a big research area. The specification could offer a guideline for specific parameters and their values based on a review of the literature. It would then offer emotional TTS developers a good baseline. Any takers???

   This "level of implementation" solution allows an unsophisticated user to get animation quickly: implementation dependant or personality dependant actions / responses. A sophisticated user could also script unambiguous actions for the tags to force the animation to follow his/her requirements. Everyone is happy except the implementer who has to do more work! Levels of implementation can help here as well.

   Related issues: what are the implied relations / semantics between tags of the sub-systems? What are the valid combinations? Which have dependencies? These also need to be specified / resolved.

2. Levels of Implementation

   As for many standards, implementation is an important issue. Languages such as Algol did not catch on because it was difficult to build compilers, whereas Pascal, a similar but simpler language, became a popular general-purpose language. It is easy to write a Pascal compiler so many people did! If VHML evolves to meet many needs, it runs the risk of becoming too cumbersome, too hard to learn, too hard to implement and hence an academic exercise at best. A similar problem was seen in the implementation of the Graphical Kernel System (GKS) [10] and the developers chose to specify the language in terms of input and output levels so that a minimal implementation could be done and certified.

   Current discussion by the VHML developers is the classifying of a VHML implementation at various levels: level 0 may mean just a visual rendering, level 1 may mean include emotions, etc. However, VHML is structured around visual and audible renderings as well as hypermedia environment awareness, so a more reasonable stratification may be along 3 or more dimensions. Implementation at $LEVEL_{ijk}$ may be used to indicate a level i implementation for visual rendering, level j for audible, etc. This would allow

implementers the ability to incrementally build a VHML system. This would mean that the VHML developers must be able to supply VHML test files for validation at these various levels. There's a history lesson here though: compare SGML and XML: who's talking about SGML today? Smaller and simpler may be better. We would welcome input on this aspect of VHML development.
Some of the following uncertainties/problems may also become part of this multi-dimensional implementation matrix.

3. Application areas and Sub-classing

It is tempting to make any new language the answer to all problems. This obviously increases its market appeal. However, doing this makes the language too general and often too large for a normal content developer to comprehend. In a note on MPEG-4 standardisation [11] indicates that the success of a standard depends in part on the following factors:

- A priori standardization
  - **Is there a real need?**
  - **Is it covered, partially covered or in the process of being covered?**
  - **Is there a danger of having a *de facto* standard?**
- Specify the minimum
  - **Take into account the generic aspects of the standard (multi-culture)**
  - **Do not overlap with already existing standards**
- One functionality – one tool
  - **Avoid options while achieving generic solutions**
  - **Too many options may lead to a failure of interoperability**

We can see from this that "incremental creep" of new tags and functionality should be resisted or at worst, controlled in some well-defined way (see "Extensibility" and "ARB" below).
VHML should be restricted to addressing the needs of directing Virtual Humans. However, if the language (or some part of it) can be used as a sub-class of some other markup language, then this is beneficial to the entire community. The fame of GKS lies in the fact that the standard was used as the basis for so many other systems that followed. This why VHML uses the Voice-XML [12] standard for speech markup which uses the Speech Synthesis Markup Language!
If this sub-classing occurs, it is important that the language be very stable, consistent and unambiguous.

4. Extensibility

Many emerging or existing standards allow for a controlled extensibility (for example, MIME types, OpenGL, X Windows) especially in areas where the application of the standard is varied. VHML already allows for the use of native language names for the tags and attributes. For example, it is possible to use the Swedish word, <arg> instead of the English word <angry> and a synonym, <joyful> instead of <happy>. An XSL stylesheet transforms the tag/attribute to the correct element and attribute names, which can then be validated by the DTD or schema. A specific stylesheet has to be constructed for each language as well as for synonyms.

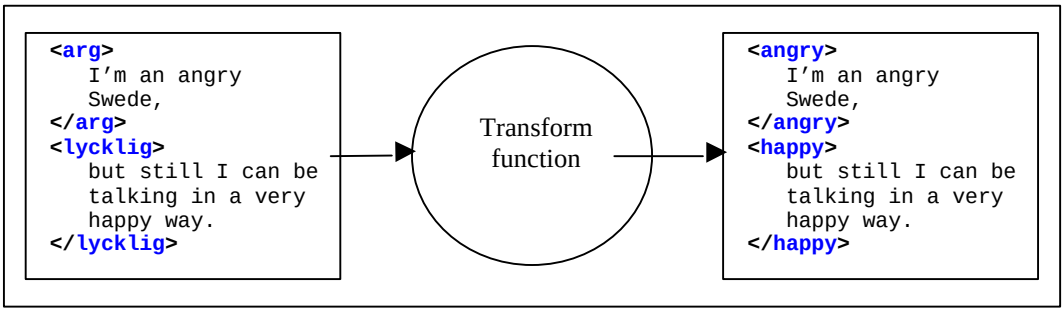

Figure 2 XSL transforming of native language tag names

We have also seen the need for extending the language to allow for exact semantic definitions. This implies that any specific tag could be re-defined to do something totally different from its **intent**. In a similar way, a tag could be defined which groups a number of existing VHML tags together as one atomic VHML tag – a macro facility. It is a small step from this to where new tags could be defined **along with their exact definition** so that the language is extended but will still be usable by all compliant implementations.

The problem is three fold, the solution moderately easy:

a) The VHML DTD must stay unchanged to accommodate the new tag
b) The new tag syntax and semantics has to be defined within the VHML framework
c) The new tag must be able to be used as naturally as possible within the text

Problems – Further Work.

1. Timing of the actions

The current specification does not address intra-action timing considerations. That is, how fast does the action start, how long to reach some level, how long at that level, etc. This attack-sustain-decay-release type timing is crucial to some applications especially those rendering subtle emotional effects. A solution could use the above simple approach or the Spline Animation of SMIL2.0 [13].

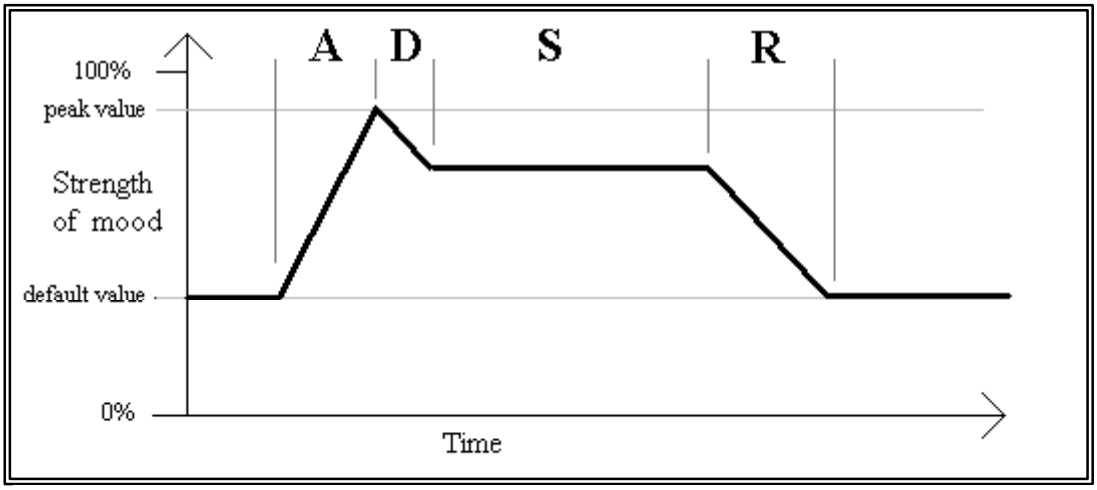

Figure 3 Attack Sustain Decay Release timing

2. Parallelism

The current specification does not address the need for concurrent rendering / directing of VHs (see [4]). This is necessary if multiple VH's are "doing" things at the same time. In keeping with the philosophy of not re-inventing the wheel, the <par>, <seq> and <excl> tags of SMIL2.0 [13] could be used. And what of tags that enable "communication" or synching between the parallel branches (e.g. virtual man speaks to virtual woman, virtual woman reacts to virtual man)?

3. Personality Specification

A Virtual Human needs a personality. Personality Theory is a major field of study in psychology used to describe individual attributes of people. It is important to note that personality is distinct from emotion. Emotions are short-term responses to an event, whereas a personality is a consistent description of the sorts of reactions a person may have. The personality of an individual will affect what emotions they perceive and display. How to model Personality?

There are a number of categories of Personality Theories (taken from [14]): *Psychoanalytical, Phenomenological, Cognitive, Trait, Behavioural, Social Cognitive, Cognitive Information Processing.* In 1963, W. Norman combined contemporary trait theories and research into the "Big Five Model" of factor-analytic study, considered a trait research benchmark. The five factors are extraversion, neuroticism, conscientiousness, agreeableness and openness. Each trait represents a range of values, for example "agreeableness" represents a range of values from agreeable to disagreeable.

Trait Name	Low Score (0.0%)	High Score (100.0%)
Conscientiousness	Unreliable, lazy, careless, negligent	Organised, reliable, neat, ambitious
Neuroticism (Emotionality)	Calm, secure, unemotional, relaxed	Worried, insecure, emotional, nervous
Openness (Culture)	Unartistic, conventional	Creative, original, curious, imaginative
Extraversion	Un-artistic, conventional	Talkative, optimistic, sociable, affectionate
Agreeableness	Rude, uncooperative, irritable	Good-natured, trusting, helpful

The Big-Five Model is trusted in the field of psychology, and is adaptable to VHML. The five factors have been found to give stable and valid results with different observers, instruments and across different adult ages. This is a strong argument for use of the Big Five model, particularly in combination with the ease in which emotions can be applied to the five personality traits. These traits can be described using natural language, making it simpler for a user to define and select a personality.
The specification of the long-term temporal dynamics of a personality (what happens when the VH gets bored, is constantly harassed, is manic-depressive) also needs to be addressed.
The specification of a personality model is not seen as part of VHML but as affecting it through the need for exact marking-up of how tags are rendered. Also, the top level <person> tag **must** include an optional personality attribute.
Current Talking Heads developed at Curtin University have rudimentary personalities of an ad hoc nature. We lean towards the Big-Five Model but we are not experts and would welcome efforts at formalisation.

4. VHML tools

There is a dearth of high-level tools for the manipulation of VHML tagged text.
[15] details a Dialogue Management Tool that lets a user create / manage VHML user dialogues.
Figure 4 and 5 show some initial work done using the widely available Gvim editor (www.vim.org) and plugins. A user can configure menus to add functionality to the editor (for example, surround highlighted text with VHML tags). The menus can be detached to provide a complex tailored markup environment. Gvim is also a folding editor that knows about syntax and colour highlights it. In Figure 5, the tags have been turned near invisible so that the actual text content can be seen.
For widespread use, a Word plugin needs to be developed with similar functionality.

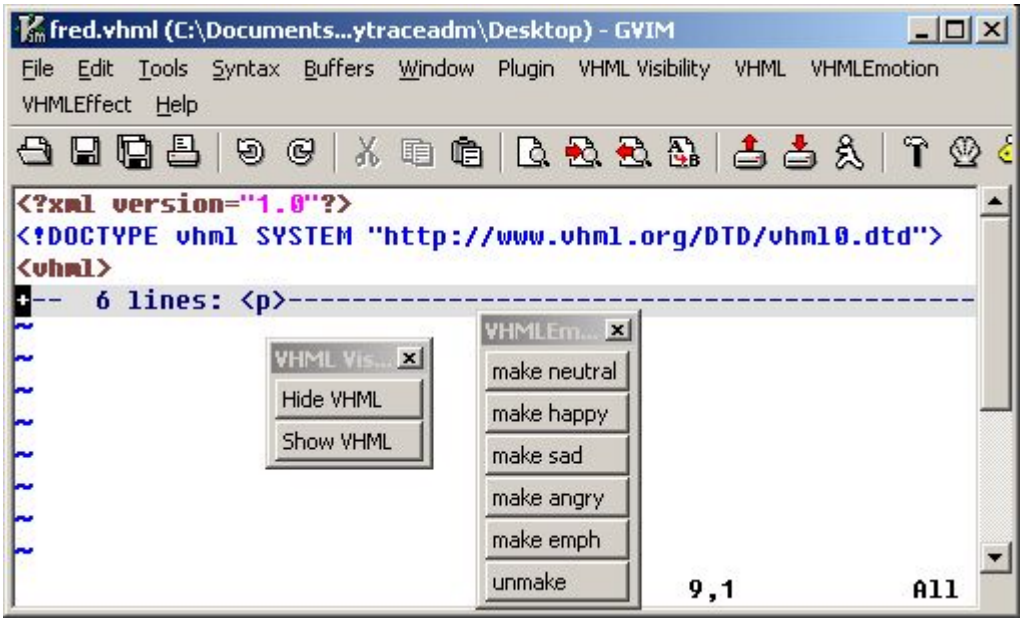

Figure 4 GVIM plugins that allow VHML markup

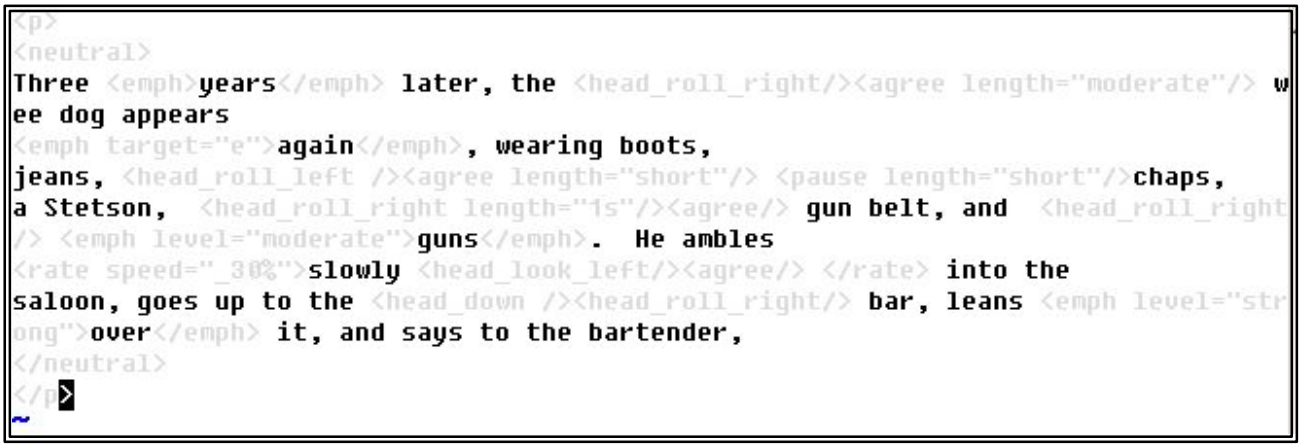

Figure 5 Seeing the forest through the trees.....

4. BAML, XHTML and DMML

BAML has been added to by one of the InterFace partners:

<left-arm-front>, <right-arm-front>,<left-arm-back>, <right-arm-back>
<left-arm-abduct>, <right-arm-abduct>,<left-arm-adduct>, <right-arm-adduct>
<left-forearm-flex>, <right-forearm-flex>
<left-forearm-turn-right>, <left-forearm-turn-left>,<right-forearm-turn-right>, <right-forearm-turn-left>
<left-hand-forward>, <right-hand-forward>, <left-hand-back>, <right-hand-back>
<left-hand-abduct>, <right-hand-abduct>, <left-hand-adduct>, <right-hand-adduct>
<left-hand-open>, <right-hand-open>
<trunk-turn-left>, <trunk-turn-right>
<trunk-front>, <trunk-back>, <trunk-bend-left>, <trunk-bend-right>
<pelvis-turn-left>, <pelvis-turn-right>, <pelvis-left>, <pelvis-right>
<left-leg-front>, <right-leg-front>, <left-leg-back>, <right-leg-back>
<left-leg-abduct>, <right-leg-abduct>, <left-leg-adduct>, <right-leg-adduct>
<left-leg-turn-right>, <left-leg-turn-left>,<right-leg-turn-right>, <right-leg-turn-left>
<left-calf-flex>, <right-calf-flex>
<left-foot-up>, <right-foot-up>, <left-foot-down>, <right-foot-down>
<left-foot-abduct>, <right-foot-abduct>, <left-foot-adduct>, <right-foot-adduct>
<left-forefoot-up>, <right-forefoot-up>, <left-forefoot-down>, <right-forefoot-down>

but as yet no higher level tags such as <walk>, <wave>, etc have been considered. Development of this is likely to take place in late 2002 / early 2003.

The XHTML sub-language is still basically the anchor tag <a> but the <meta> tag is also seen as useful as is the XHTML <embed> tag especially given the ubiquitousness of the Web and the need for an application to link to other data. For example, simply using

```
<meta http-equiv="refresh" content="1; URL=hard-data-here.html">
```

can load up the “hard-data-here.html” page.
An argument does exist for including the content **in** the VHML document but this would be at the expense of full XHTML.

DMML has developed in an ad hoc fashion but needs to be formalised and is currently under discussion.

It has been suggested that a Hand Gesture Markup Language is necessary since many cultures rely on extensive hand motions to illustrate, punctuate and augment their conversations.

Architectural Review Board

VHML owes its existence to the work done primarily by John Stallo. Quoc Huynh and Simon Beard also contributed to its development. The InterFace project has been discussing VHML since the end of 2000 and in February 2001 an initial rudimentary specification was tabled at one of the project meetings. Since then, the specification has been reviewed and sub-language implementations have been evaluated. In the second half of 2001 Emma Wiknertz, Linda Strindlund and Camilla Gustavsson made the language more solid, homogenous and complete – they made the specification a quality document. Simon Beard has used this specification to create the first near complete implementation [4]. Since its birth, Andrew Marriott has provided a guiding hand and been a sounding board for its development.

We have seen from this paper that VHML is still evolving and being corrected. But what if someone grabs the specification and starts adding things to it that “we” do not want. Changing the DTD is easy, adding tags is easy, changing the intent of tags is easy. What if someone adds an entire sub-language? If done well this is not a problem: but what if two groups develop a language targeted at the same area such as Hand Gestures? Who resolves the conflicts? Whose language gets added to VHML (assuming that people are interested in doing this)? If this cannot be resolved, then two new standards will emerge and this is counter-productive.

And who is the “we” that has control? A closed proprietary standard is not something that is useful to researchers or industry. An open standard is effective because people know that they can affect its development. If I say that VHML is copyright and that the EU says you can only play with it under our rules, then you will probably tell me to go and stick my head in a bushfire (wise old Australian saying). VHML **is** copyright (as is any work that has been published) but it is also open to input, discussion and change.

It is planned that an Architectural Review Board (ARB) for VHML be set up to guide its future development. In this way input can be moderated, conflicts resolved, etc by those with knowledge and an invested interest. By invested interest, I mean those who put time, effort and knowledge into serious development as well as those who have a legitimate financial interest in the language. If VHML is not seen as commercially viable, then it will remain just an academic tool.

Conclusion

A Talking Head directed by text with markup tags is perceived by users as being more human-like than one that just speaks [1]. Using a markup language may mean that a Virtual Lecturer is seen as being erudite and approachable, a Virtual SalesPerson as trustworthy and helpful, etc. VHML is one such markup language.

Acceptance of VHML by researchers and users will depend upon it meeting the criteria required of any potential standard:

- Completeness – is it complete in itself for the targeted audience?
- Simplicity – is it simple enough for people to learn and quickly use?
- Consistency – is it consistent so that people can use it without needing a complex manual?
- Intuitive – is it obvious what the tags do and mean?
- Unambiguous – is it unquestionably spelt out as to what the language does?
- Abstraction – does it provide the right level of abstraction for the targeted audience and function?
- Extensible – is it possible to extend the language without breaking it?
- Usability – is it usable in terms of implementation and functionality?
- Effective – does evaluation indicate that the user experience is positive and/or rewarding?

VHML's evolutionary development has been influenced by many researchers from many disciplines. It will continue to be defined / refined / corrected through the ARB with input from workshops, discussion groups and researchers. Those interested in discussing any of the previous uncertainties or problems are encouraged to contact the authors.

References

[1] A. Marriott, S. Beard, H. Haddad, R. Pockaj, J. Stallo, Q. Hyunh, and B. Tschirren, "The Face of the Future," *Journal of Research and Practice in Information technology*, vol. 32, pp. 231-245, 2000.

[2] A. Marriott, S. Beard, J. Stallo, and Q. Huynh, "VHML - Directing a Talking Head," in *Active Media Technology*, vol. LNCS 2252, J. Liu, P. C. Yuen, C.-H. Li, J. Ng, and T. Ishida, Eds. Hong Kong: Springer, 2001e.

[3] A. Marriott, "A Facial Animation case study for HCI: the VHML-based Mentor System," in *MPEG-4 Facial Animation - The standard, implementations and applications*, I. Pandzic and R. Forchheimer, Eds. New York: John Wiley, 2002.

[4] S. Beard, D. Reid, and A. Marriott, "MetaFace and VHML: A First Implementation of the Virtual Human Markup Language," Curtin University of Technology, Perth, Australia, Workshop paper - to be published 2002.

[5] JRPIT, "Experiment details for "The Face of the Future". Online at http://www.interface.computing.edu.au/papers/jrpit-hci/html/ and http://www.interface.computing.edu.au/papers/jrpit-hci/word-docs/," 2001b.

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

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

[8] ISO/IEC, "Text for ISO/IEC FDIS 14496-2 Visual. Online at http://www.cselt.it/mpeg/working_documents.htm," ISO/IEC 14496-2, Nov 1998 1998b.

[9] I. R. Murray and J. L. Arnott, "Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion," *Journal of the Acoustical Society of America*, vol. 2, pp. 1097-1108, 1993.

[10] F. Hopgood, D. Duce, J. Gallop, and D. Sutcliffe, *Introduction to the Graphical Kernel System (GKS)*: Academic Press, 1986.

[11] F. Marques, "Mpeg-4 Standardisation Criteria." UPC, Barcelona, Spain, 2001.

[12] VoiceXML, "VoiceXML Forum, www.voicexml.org," 2000.

[13] SMIL2.0, "Synchronized Multimedia Integration Language (SMIL 2.0)," 7 August 2001.

[14] L. A. Pervin, *Personality Theory and Research, Fifth Edition*. New York: John Wiley & Sons, 1989.

[15] C. Gustavsson, L. Strindlund, and E. Wiknertz, "Dialogue Management Tool," presented at The Talking Head Technology Workshop of OZCHI2001, The Annual Conference for the Computer-Human Interaction Special Interest Group (CHISIG) of the Ergonomics Society of Australia. http://www.ep.liu.se/exjobb/isy/2002/3188/, Fremantle, Australia, 2001.