生命的エージェントによる感性的マルチモーダルコンテンツ記述と生成

Description and Generation of Affective Multimodal Contents using Lifelike Agents

東京大学大学院情報理工学系研究科
School of Information Science and Technology, University of Tokyo

石 塚　満
Mitsuru Ishizuka

<研究協力者>

東京大学大学院工学系研究科
School of Engineering
University of Tokyo

東京大学大学院情報理工学系研究科
School of Information Science and Technology
University of Tokyo

土肥　浩
Hiroshi Dohi

Helmut Prendinger,　Santi Saeyor

Abstract: As an important application area of new TTS(text-to-speech) with emotions, we have conducted our research on a synthesis tool for mutlimodal content production systems using lifelike agents. We have developed a family of XML-based content description languages called MPML (Multimodal Presentation Markup Language) for controlling lifelike agents in various environments including the Web, 3D space (i.e., VRML space) and mobile phones. Emotion, together with personality, is an important factor for improving the lifelikeness and believability of the agents, and thus making the presentation attractive, impressive and memorable. Accordingly, MPML provides a tag for scripting emotional behaviors and voice. It is cumbersome, however, for a content author to choose and write down the appropriate tag description of emotion everywhere necessary in the possible flow of a scenario. An artificial emotion module called SCREAM allows to design agents that autonomously generate emotionally and socially appropriate behaviors based on their character profile. Combined with MPML, SCREAM facilitates a high-level scripting.

Keywords: lifelike agent, multimodal interface, description language, emotion, TTS

1．まえがき

顔と姿を持ち音声機能を有する生命的エージェント（lifelike agents；擬人化エージェントやECA(embodied conversational agents)などとも呼ばれる）を用いるマルチモーダルインタフェースやコンテンツが出現し始め，複雑化が進行する情報化社会の中で理解しやすく親しみやすい新形態のマルチモーダルメディアとして発展が期待されている[Cassell 00, Prendinger 04a]．いくつかの先行的研究開発により，その効果も実証されるようになってきている．多くの技術要素の集積が必要なため，以前は実現が容易ではなかったが，数年前よりツールとして提供される要素技術が出現し始め，試行的使用は容易になりつつある．一方で，自由度が高く感性的で魅力的なコンテンツやインタフェースの作成には課題も多く，必ずしも容易ではない．

本研究では高度音声合成の応用として，この生命的エージェントに関しコンテンツ記述言語MPML(Multimodal Presentation Markup

Language)と感性機能を中心にして研究を実施し，成果を得ている．

生命的エージェントの認知的意義や効用については，以下のような知見が基礎として存在する．

- 1960 年代の A. Mehrabian によるノンバーバルコミュニケーションの役割の重要性についての研究（人間間のコミュニケーションで表情やジェスチャ等によるノンバーバル情報は 55%，イントネーションや音質による部分は 38%もの情報伝達を担っている．）
- 1990 年代になってからの B. Reeves & C. Nass の”Media Equation”の考え方．（”Media = Real Life”，即ち，人間は接するメディアを人工物としてでなく，生命体として認知する性向を遺伝的に有する．）
- Persona Effect（学習システム等において生命的エージェントの存在は生徒等にポジティブな効果を与える．直接学習が効率的になる訳ではないが，意欲を高めたり，興味を喚起したりする効果を持つ．）

この生命的エージェントを我々の情報空間の真のパートナとなるように育てることに向けて，幾つかの観点からの研究開発が要請されるのであるが，ここでは我々の記述言語 MPML と感性的機能を中心に報告する．

エージェントのプレゼンテーションは平板なものになりがちだが，感情表現の付加はエージェントの生命感，信頼感を向上させる上で重要である．エージェントの感情は視聴者の感情も呼び起こし，親近感，エンタテイメント性，モチベーション等を向上させる効果を有する．感情は喜び（幸福感），悲しみ，驚き，怒り，嫌悪，恐れ・・・などの言葉で語られるが，場合によりそのカテゴリ分けは不統一で，根拠も不十分なものであった．MPML では最も包括的な OCC モデルによる感情を扱うようにしており，感情状態をタグで囲んで記述する．この時，発話の前後に感情による動作をし，感情に応じて TTS の発話スピード，ピッチ，ピッチの変動幅，強度の音声パラメータを変化させるようにしている．これにより少ない記述で，音声を含めて感性的なエージェントを生成できるようにしている．

2．記述言語 MPML

コンピュータゲームコンテンツに見られるように資金とプロのクリエータの労力によれば，感性的にも優れたコンテンツを作成できるレベルの技術は存在し，周知のようにこの面で我が国は世界をリードしている．しかし，一般の人々がこのようなレベルのコンテンツを作成できるという訳ではない．

Web 上のコンテンツの急速な拡大に見られるように，マルチモーダルコンテンツの普及，浸透を図る上でも誰でもが容易に作成し視聴できるような環境，及び標準的なコンテンツ記述法を整えることが重要となる．キャラクタエージェントを用いるコンテンツの記述法として，このような方向を目指し，世界で XML-based のマークアップ型記述言語の研究開発が進められている．我々の MPML (Multimodal Presentation Markup Language)もその一つである．（他には，VHML, CML/AML, APML, RRL-NECA, BEAT などがある．）標準化は望ましく必要なのであるが，使用するキャラクタエージェント・システムの違い，どのレベルのコンテンツ・オーサを対象とするのか（プロのクリエータレベルにも対応か，一般の人々向けか），これらに関係してどの粒度レベルのコントロールを行うか，あるいは許容するか等の点において合意が取れず，進展は捗々しくない．

MPML [MPML Homepage, 筒井 00, Descamps 01a, 01b, Prendinger 04b] は Web における HTML のように，一般の人々の誰でもが

```
<mpml>
<head>
<title> MPML Presentation </title>
</head>
<body>
<page id='first' ref='self_intro.html'>
<emotion type='happy-for'>
<speak>
   I am Mitsu Ishizuka from the Univ. of Tokyo.</speak>
</emotion>
</page>
</body>
</mpml>
```

図1 MPMLの記述例

容易にキャラクタエージェントを用いるマルチモーダルコンテンツを記述できるようにすることを主要な狙いとしている．”Anytime”, “Anyplace”, “Anyone”をスローガンにしているが，このAnyone は「誰でもがオーサリングできる」という狙いを意味している．

Powerpoint 等によるプレゼンテーションは，図表も含め視覚的でかつ論理的に整理されたプレゼン資料と，人間のプレゼンタの表情や身振りも含めた音声によるマルチモーダルな情報提示，伝達であり，人間の認知的受容性に適合し，現代の主流の形態になっている．しかし，同時刻，同じ場所にプレゼンタが存在しなければならないという大きな制約がある．MPML の Anytime, Anyplace は，人間のプレゼンタの役割をキャラクタエージェントに代行させ，ネットワークを介してプレゼンテーションコンテンツを伝達できるようにし，時間と位置の制約から自由になるということを意味している．Anyplace については，最近では携帯電話向きの MPML-Mobile 版[Santi 03]の開発も行っているので，モバイル環境への拡大も意味するようになってきている．

記述の詳細は省くが，図１のようなマークアップタグによるXML 言語であり，VB ScriptやＪava Script プログラムのようにプログラミング言語を知らなくても，HTML を記述できる人なら新たな20～30 程のタグの使い方を知ることにより記述可能である．（HTML のように MPML の Graphical Editor の初期版も用意されているので，MPML 自体をたとえ知らなくても記述することは可能になっている．）ビデオや音声データも含むメディア同期用にSMIL の基本機能を含んでいる．

PC 上でのプレゼンテーションコンテンツの場合，背景はHTML の Web ページとして作成することになる．XML 準拠の言語仕様とすることのメリットは，XML 対応のブラウザを使用できたり，関係のツール類を利用できることがある．MPML から実際のプレゼンテーションを駆動する言語レベル（Java Script など）への変換はプログラムを使う版もあるが，XSLT で実装するとユーザはこの変換を意識することなく，普通の Web コンテンツと同様に Web ブラウザ上でクリックするだけで MPML コンテンツが視聴できる．

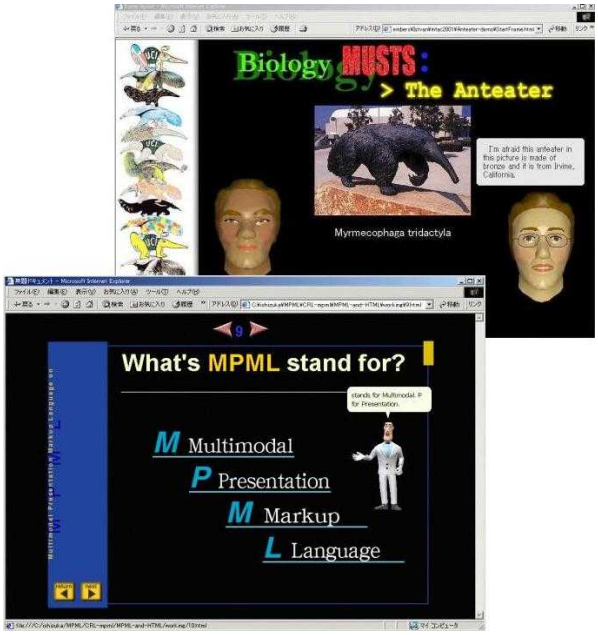

図２ MPML の画面例

図３ MPML-VR の画面例（３次元）

図４ MPML-Mobile の画面例

キャラクタエージェントとしては MS Agents を基本的にサポートしているが，ドライバ部分のプログラムを書くことにより各種エージェントを使うことができ，実際，3D VRML 空間での H-Anim 規格のエージェント，携帯電話(i-mode, KDDI-au, Vodafone)上のエージェントにも対応するようになっている．後述する顔表情豊かな SmArt Agent にも対応している．

図２は MPML による画面のスナップショットを例示している．図３は 3D VRML 空間に対応する MPML-VR 版の画面例である．図４は携帯電話に対応する MPML-Mobile 版の例である．

3．エージェントの感情

感性的なコンテンツを誰でもが容易に作成しやすくする環境を整えていくことは今後の重要な課題であるが，感情表現はその一つの重要な要素である．

エージェントのプレゼンテーションは平板なものになりがちだが，感情表現の付加はエージェントの生命感，信頼感を向上させる上で重要である．エージェントの感情は視聴者の感情も呼び起こし，親近感，エンタテイメント性，モチベーション等を向上させる効果を有する．

感情は喜び（幸福感），悲しみ，驚き，怒り，嫌悪，恐れ・・・などの言葉で語られるが，場合によりそのカテゴリ分けは不統一で，根拠も不十分なものであった．これを整理し，1988 年に著書として発表された認知評価理論 (cognitive appraisal theory)，あるいは提唱者３名の名前を取り OCC モデル[Ortony 88]と称される感情モデルでは，最も包括的な 22 種の感情が用いられる．MPML では，後述の人工感情モジュールとの関係もあり，この OCC モデルによる感情を扱うようにしており，ユーザの感情センサとの対応からより簡単な Valence-Arousal の２軸による感情モデルも併用している．

MPML では感情表現の指定は直接的には図 1 に示されるようにタグで囲んで記述する．この時，発話の前後に感情による動作をし，表１に例示するような感情により発話スピード，ピッチ，ピッチの変動幅，強度の音声パラメータを変化させるようにしている．既存の TTS エンジンで提供されている範囲の音声パラメータ制御であり，今後の改良が期待されるところではあるが，ここでは単に音声だけでなく動作（ジェスチャ）も付随するので，不十分さが補われている．

すべての感情指定タグをマニュアル入力するの

表１ 感情と音声パラメータ

Emotion	Fear	Anger	Sadness	Happiness	Disgust
Speech rate	much faster	slightly faster	slightly slower	faster or slower	very much slower
Pitch average	very much higher	very much higher	slightly lower	much higher	very much lower
Pitch range	much wider	much wider	slightly narrower	much wider	slightly wider
Intensity	normal	higher	lower	higher	lower
Pitch changes	normal	abrupt on stressed syllables	downward inflections	smooth upward inflections	wide downward terminal inflections

Emotion	Fear	Anger	Sadness	Happiness	Disgust
Speech rate	+30	+10	-10	+20/-20	-40
Average pitch	+40	+40	-10	+30	-40
Loudness	—	+6	-2	+3	—

は煩雑となる．そこで，エージェントの人工感情モジュールの役割を果たす SCREAM(SCRipting Emotion-based Agent Minds)を開発している[Prendinger 01, 02a, 02b]．これは OCC モデルに準拠してエージェントの感情を計算するモジュールである．OCC モデルでは，1)事象の結果(consequence of events)に対する感情（これを更に他者に対する結果と自己に対する結果の感情に分ける），2)エージェントの行動(action of agents)に関する感情（このエージェントを更に自己と他者に分ける），3)物への心胆(aspect of objects)の感情に分けているが，SCREAM はこれらをルールベースにより決定するようになっている．更に上司や部下といった社会的関係により，感情の表出を制御する Social Filtering 機能を実装している．図５は SCREAM の構成と実装の図を示している．

MPML からは consult タグにより外部モジュールである SCREAM を呼び出し，エージェントの観桜を決定するようにしている．

MS Agents などは顔の表情は豊かではないので，豊かな顔表情表出が可能な独自のエージェント SmArt [Barakonyi 01]も作成している．

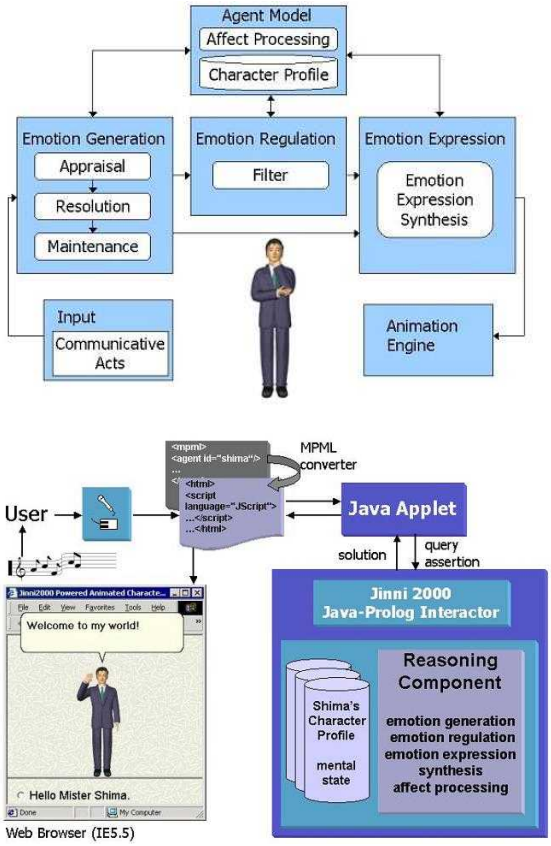

図５ SCREAM の構成と実装

4．その他の機能

先に例示したように，MPML は３次元 VRML 空間，携帯電話（DoCoMo i-mode, KDDI-AU, Vodafone)への対応版も開発している．

実用性重視のためコンテンツはスクリプト既述を主としているが，自律性の拡大も図っている．例えば，Chatbot と組み合わせて，想定外の自由対話にも対処できるようにし，伝達しようとするコンテンツとのスムースな切り換えを可能にしている[Mori 03]．

5．むすび

高度音声合成の応用として有望な，生命的エージェントを用いるコンテンツやインタフェースの記述言語と，その関連技術について報告した．エージェントの生命感，信頼性を向上させるためには，感情の表出が重要な要因となる．現状では感情を伴う音声は，既存の TTS のパラメータ制御で生成しているが，単に音声だけでなく動作も付随するので，不十分な点が補われている面がある．感情をもつ高度音声合成技術が一般に使用できるようになると更に望ましいことになる．

参考文献

[Barakonyi 01] I. Barakonyi and M. Ishizuka: A 3D Agent with Synthetic Face and Semiautonomous Behavior for Multimodal Presentations, Proc. Multimedia Tech. And Application Conf. (MTAC2001), pp.21-25, Irvine, California (2001.11)

[Cassell 00] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds.): Embodied Conversational Agents, The MIT Press (2000)

[Descamps 01a] S. Descamps, H. Prendinger and M. Ishizuka: A Multimodal Presentation Markup Language for Enhanced Affective Presentation, Advances in Education Technologies: Multimedia, WWW and Distant Education (Proc. Int'l Conf. On Intelligent Multimedia and Distant Learning (ICIMADE-01), Fargo, North Dakota, pp.9-16 (2001.6)

[Descamps 01b] S. Descamps, I. Barakonyi and M. Ishizuka: Making the Web Emotional: Authering Multimodal Presentations using a Synthetic 3D Agent, Proc. OZCHI-2001, pp.25-30, Perth, Australia (2001.11)

[Mori 03] K. Mori, A. Jatowt and M. Ishizuka: Enhancing Conversational Flexibility in Multimodal Interactions with Embodied Lifelike Agents, Proc. Int'l Conf. On Intelligent User Interfaces (IUI2003), pp.270-272, Miami, Florida (2003.1)

[MPML Homepage] http://www.miv.t.u-tokyo.ac.jp/MPML/

[岡崎 02]岡崎，S. Saeyor，土肥，石塚：マルチモーダルプレゼンテーション記述言語 MPML の 3 次元 VRML 空間への拡張，電子通信学会論文誌，Vol. J85-D, No.9, pp.915-926 (2002.9)

[Ortony 88] A. Ortony, G. L. Clore and A. Collins: The Cognitive Structure of Emotion, Cambridge Univ. Press (1988)

[Prendinger 01] H. Prendiger and M. Ishizuka: Let's Talk! Socially Intelligent Agents for Language Conversation Training, IEEE Trans. On System, Man and Cybernetics, Part A, Vol.31, Issue 5, pp.465-471 (2001.9)

[Prendinger 02a] H. Prendinger and M. Ishizuka: SCREAM: Scripting Emotion-based Agent Minds, Proc. 1st Int'l Joint Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS-02), Bologna, Italy, pp.350-351 (2002)

[Prendinger 02b] H. Prendinger, S. Descamps and M. Ishizuka: Scripting Affective Communication with Life-like Characters in Web-based Interaction Systems, Applied Artificial Intelligence, Vol.16, No.7-8, pp.519-553 (2002)

[Prendinger 04a] H. Prendinger and M. Ishizuka: Life-like Characters—Tools, Affective Functions and Applications, Springer-Verlag (2004)

[Prendinger 04b] H. Prendinger, S. Descamps and M. Ishizuka: MPML: A Markup Language for Controlling the Behavior of Life-like Characters, Jour. of Visual Language and Computing, to appear, Vol.15, No.2, pp. 83-203 (2004.4)

[Santi 03] S. Saeyor, K. Uchiyama and M. Ishizuka: Multimodal Presentation Markup Language on Mobile Phones, AAMAS Workshop Proc. (W10)-Embodied Conversational Characters as Individuals, Melbourne, Australia, pp.68-71 (2003.7)

[筒井 00]筒井，石塚：キャラクタエージェント制御機能を有するマルチモーダルプレゼンテーション記述言語，情報処理学会論文誌，Vol.41, No.7, pp.1976-1986 (2000.7)