マルチモーダル対話記述言語に関する標準化動向

新田 恒雄

*豊橋技術科学大学　大学院工学研究科

〒441-8580　愛知県豊橋市天伯町雲雀ヶ丘１－１
Email:　nitta@tutkie.tut.ac.jp

あらまし：マルチモーダル対話への関心が，次世代携帯端末や様々な情報端末を中心に高まっている。本報告では，マルチモーダル対話（MMI）記述言語に関する動向を，W3C-MMI-WG 活動を中心に紹介すると共に，MMI 記述言語の現状比較，および W3C の他の標準化との連携について解説する。

キーワード：　マルチモーダル対話，標準化，W3C，XML，SMIL，VoiceXML，SALT，XISL

Standardization of Multi-modal Interaction Description Language

Tsuneo　NITTA

Graduate School of Technology, Toyohashi Univ. of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi 441-8580, JAPAN
Email: nitta@tutkie.tut.ac.jp

Abstract :　As the multimedia infrastructure is progressively developed, the multi-modal interaction (MMI) used in 3GPP and various MM applications is becoming increasingly important. In this paper, standardization of MMI description language, especially W3C-MMI Working Group activities is introduced. Moreover, currently proposed MMI description languages and MMI-related W3C standardization are also described.

Key words :　Multimodal Interaction, Standardization, W3C, XML, SMIL, VoiceXML, SALT, XISL

1. はじめに

マルチモーダルインタラクション（MMI）の標準化に関する議論が盛んに行われるようになった。W3C では MMI ワーキンググループ (W3C-MMI-WG [1])が結成され，記述言語の仕様策定が進められている。また並行して様々な研究機関で，MMI の記述法に関する検討がなされている[2], [3], [4], [5], [6]。以下では，著者も一員となっている W3C-MMI-WG の活動を紹介すると共に，MMI 記述言語の現状比較，および W3C の他の標準化との連携について解説する。

2. W3C-MMI-WGの活動

VoiceXML の仕様が固まると共に([7]; VoiceXML は 2.0 の策定をほぼ終えているが，パテントの扱いに関して RF (Royalty Free) に戻すようにとの W3C Director 裁定が 6 月に出され，RAND (Reasonable and Non-disclosure) を前提としてきた各機関に対する調査が始められるところである)，次世代携帯 (3GPP: The 3rd Generation Partnership Project)，PDA，kiosk 端末，カーナビ等を念頭においた MMI 記述言語の策定活動が，本年 3 月ニースにおける W3C technical plenary

から始められた [1]。これまでに行われた電話会議と，今回ボストンで開催された二回目の Face-to-Face (F2F) に至る概要を以下に紹介する（参加者38名。日本からはNECと著者のTUT。次回は，9月にフィンランドの Tempere において Nokia ホストで開催予定）。

MMI-WG による標準化は，W3C の他の勧告策定過程と同じく，表1に示した日程で進められる（年月日はあくまで予定である）。現在は，年末の WD 全体構成の決定を目指して，要求される機能 (Requirement) と仕様 (Specification) をユースケースを参考に詰める作業を行っている。以下に現在行われている作業概要を述べる。

表1　W3C-MMI-WG のマイルストーン（予定）

Working Draft : WD-1　(全体構成)	2002-12
: WD-2	2003- 05
: Last call	2003-08
Candidate Recommendation(勧告候補):	2004-02
＊　charter calls を 2004 年の初めに予定	
Proposed Recommendation (勧告案)	2004-03
Recommendation (勧告)	2004-05

2.1 応用システムとその分析

参加機関から提供された MMI の use case を XML ドキュメントとしてまとめた。今後，この中から代表的な例を取り出し，エンドユーザとプロバイダ間のやり取りと問題点，そして解決法を詳細に分析する。

2.2　要求される機能 (Requirement)

現時点の要求を以下に示す（変わる可能性有り）。

(1) 一般仕様：

― 機器性能の違いをサポート (scalability)。

― supplementary use（他のモダリティでも同じ対話ができる）と complementary use（モダリティにより異なる対話が組合せられる）をサポート。

― モダリティ間はシームレスに同期。

このほか，ユーザの端末環境に関するプロファイルの最適化・適応といった機能も必須(must)ではないが入れられている。

(2) 入力モダリティ：

― オーディオ, キー, ポインティング機器, ペン, ゲームコントローラ等をサポート。

― 入力に対する処理の指示方法をサポート。

― mutually exclusive modes（一度には一つのモダリティ）/ simultaneous input（同時的入力）/ composite input（複数の異なるモダリティを一つに構成）をサポート。今後さらに議論。

― 新しい機器・モダリティの付加をサポート。

このほか, time stamping も必須ではないが入れられている(should)。一方，gaze 認識，audio-visual 情報を用いた認識は，当面必要性が低いとされた (nice to specify)。

(3) 出力メディア：

― オーディオ（prompt, playback 含む），ビジュアル(XHTML, SVG)，SMIL オブジェクト，streaming をサポート。

― 出力に対する処理の指示方法をサポート。

― sequential/ simultaneous media output をサポート。各メディアの粒度も変えられる。

このほか，time stamping，二つ以上の window への表示は今回除かれた。

(4) アーキテクチャ，統合・同期に関して：

― SMIL の利用

― XHTMLmodularization として表現可能

― Xforms との共存

― 利用可能なモダリティの検出

― 様々なレベルでの同期の粒度

― 入出力に対する分散処理

2.3　仕様　(Specification)

ここでは，これから MMI アーキテクチャ（components, events, execution model），MMI ドキュメント（XHTML コンテナ），および MMI input/ output の仕様が，幾つかの use case に対する分析を基に話し合われる。

2.4　並行作業

仕様策定のために，W3C MMI Framework の提示（アーキテクチャ），イベント処理法，さらに NLSML (Natural Language Semantic Markup Language), Pen Input に関する検討グループが発足している。NLSML は入力モダリティ間の統合も扱うことになる。

3. MMI記述言語の現状比較

W3C-MMI-WG の目指す言語は，今後，**2.** および次の**4.** で説明する仕様条件を満たすことになる。現時点では，周辺の標準化作業とその実装

の問題から，条件を満たす言語はないが，以下にこれまで提案されたアプローチを著者らの提案しているXISLと比較しながら紹介する。

3.1 XISL

XISLは，XMLコンテンツに対するMMIを記述する言語である[8], [9]。XISLでは，ある目的を持った1組みの対話をdialogで記述し，複数のdialogにより対話シナリオを構成する。各dialogは対話の最小単位を表すexchangeから構成されている。また，各exchangeはoperation (入力記述部)とaction (動作記述部)をそれぞれ一つずつ持つ。
単一の入出力モダリティは，operation内部のinputとaction内部のoutputにより表される。inputにはシステムが受け付けるユーザからの入力を，outputにはそれに対応する動作を記述する。また，operationやaction内では，マルチモーダルな入出力を制御するタグが用意されている。これらのタグによって，逐次的(sequential)入出力，同時並行的(parallel)入出力，および択一的(alternative)入力を記述することができる。さらに，対話遷移のために<call>や<goto>を用意し，複雑な対話制御を可能にしている。

XISLでは，inputおよびoutputタグの属性と内容の種類のみを規定し，属性値の記法や，内容の詳細についてはXISLの仕様外としている。このことにより，モダリティの拡張や修正があった場合にも，XISLは仕様変更なしに多様な端末環境に対応することができる。

3.2 SMIL

SMIL[10]はマルチメディア出力のストリーミングを制御するための言語であり，動画出力を行う各種ソフトウェアで実際に利用されている。またSMILにイベント記述言語ReXを組み合わせることにより，マルチモーダル記述言語として利用する試み[3]もある。SMILは出力の制御に関して優れた記述力を持っており，XISLでも記述可能な同時的，逐次的，択一的な出力に加えて，各種イベントを契機とした出力や，時間制御（例えば，アニメーションを出力開始して5秒後に音声出力を始めるなど）が記述可能である。

一方，SMILは出力に関しては優れた記述力を持っている半面，対話の制御や情報管理に関して記述力が不足している。XISLでは対話の階層構造により，各種レベルの変数や対話を設定できるが，SMILではそのような設定や変数などの情報管理が困難である。こうしたの特徴から，MMIのシナリオに関してはXISLの方が容易に記述できる。

3.3 SALT

SALT[4]はXHTML文書等に対して音声インタフェースを付加するためのタグセットであり，webページ上での音声によるフォームの埋め込みや，音声によるテキストの読み上げを可能にする。最大の特徴は，現状の技術，特にXHTMLとの親和性が高いことである。SALTのタグはXHTML文書に埋め込む形で記述され，音声認識の文法をXHTMLの<input>タグにバインドすることにより，音声によるフォーム入力を可能にしている。XISLの<input>や<output>においても，XHTMLの技術を直接利用して属性や内容を記述する手段を提供することが必要と考えている。

一方，SALTでは対話制御の記述ができず，制御はSMILもしくはスクリプトで記述することになる。また，XHTMLにSALTを埋め込む方法では，コンテンツとインタラクションが同一文書に混在する。さらにSALTではモダリティとして主に音声の追加を想定している。これに対して，XISLはXISL自身で対話を制御し，インタラクションとコンテンツを分離すると共に，最初から多様なモダリティの利用を想定している点が異なる。

3.4 VoiceXML+XHTML

VoiceXML[7]は電話による音声対話を対象とした対話記述言語であり，ユーザの発話文法や対話の制御など，音声によるwebサービスで必要な様々な事項を記述できる。また対話の階層構造など，対話制御に関して優れた記述力を持つ部分があり，XISLでもこれらの特徴を取り入れている[8]。VoiceXMLは電話による音声対話のみを対象としているが，これに画面操作を融合させるためのアプローチも検討されつつある[5], [6]。

しかしながら，これらのアプローチではSALTと同様，音声とポインティング等，限られたモダリティを想定しているのに対し，XISLではモダリティの拡張性を考慮し，モダリティの記述に自由度を持たせている点が異なる。またVoiceXMLでは，スロットフィリング（ユーザが項目を埋めていく）の考えに基づいた対話制御を行っており，同時的，逐次的，択一的なシナリオの進行を記述するための明示的な方法がないため，対話制御に関してはXISLの方が記述しやすい。

4. W3Cの他の標準との連携について

Web 上のアプリケーションに対する MMI は，それがベストの解決法であるかは別に XForms を中心に形成される。XForms モデルは，XHTML のフォームを含む多様な UI の選択と，デバイス非依存 (DI: Device Independence) の実現を目標に策定が進められている[11]。フォームは XML インスタンスデータを外部から収集し（図 1 参照），制約をバインドして処理する。MMI の入出力は，以下に概要を説明するイベント処理機構を通して行われることになる。

DOM (Document Object Model) Level2 では，XML パーザを通して得た DOM ツリー上にイベントの流れを記述し，ノードにイベントハンドラを登録するイベントシステムが加えられている（[12]；イベントにイベントハンドラをバインドする記述方法に関しては XML Events [13] で定義)。イベントが登録されたノードは Event Target と呼ばれ，この特定のノードがイベントを取得すると，Event Listener が呼び出され，イベント処理が行われる。ツリー上のイベントの流れには，子から親への流れ (Event Bubbling) と，親から子への流れ (Event Capture) がある（図 2 参照)。イベントの流れは，Event Listener がイベントをキャンセルすることで伝播を終了する。

MMI では複数の入出力から（非同期に）分散してイベントが生じる。これらの課題に対処する方法が，現在，既存の XML Events, XML Protocol, SIP (Session Initiation Protocol) 等をベースに検討されている。

このほかの関連する技術として，端末の能力やユーザの好みにより，コンテンツの提供の仕方を変える機構として，HTML のヘッダ情報を利用する方法よりも柔軟性の高い，CC/PP (Composite Capability / Preference Profiles) [14] がある。

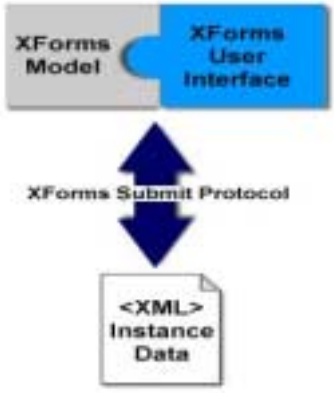

図 1　XForms モデルと UI，Data

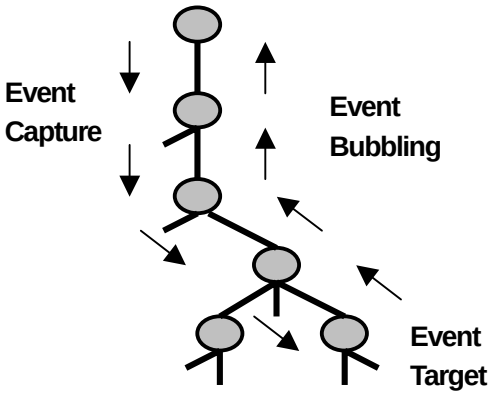

図 2　DOM ツリー上のイベントの流れ

5. お わ り に

W3C の MMI 記述言語策定活動を中心に，領域の動向を述べた。今後，多くの研究者が国際標準化活動に関心を向けられ，貢献されることを望む。

参 考 文 献

[1] http://www.w3.org/2002/mmi/

[2] T.Nitta, K.Katsurada, H.Yamada, Y. Nakamura, S.Kobayashi: XISL: An Attempt to Separate Multi- modal Interaction from XML Contents, Proc. of EUROSPEECH 2001, pp.1197-1200(2001).

[3] J.L.Beckham, G.D.Fabbrizio, N.Klarlund: Towards SMIL as a Foundation for Multimodal, Multimedia Applications, Proc. of EUROSPEECH 2001, pp.1363-1366(2001).

[4] http://www.saltforum.org/

[5] http://www.w3.org/TR/xhtml+voice/

[6] 植田，秋田，荒木，西本，新美： VoiceXML のマルチモーダル化の検討，情処研報告 2001-SLP-38，pp.43-48(2001).

[7] http://www.w3.org/TR/voicexml/

[8] 桂田，大谷，中村，小林，山田，新田：“多用な端末からのアクセスを可能にする MMI アーキテクチャ”，情処研究報告 2002-SLP-40，pp.51-56 (2002)

[9] http://www.vox.tutkie.tut.ac.jp/XISL/XISL.html

[10] http://www.w3.org/AudioVideo/

[11] http://www.w3.org/TR/xforms/

[12] http://www.w3.org/TR/DOM-Level-2-Events/

[13] http://www.w3.org/TR/xml-events

[14] http://www.w3.org/Mobile/CCPP/