# VoiceXMLのマルチモーダル化の検討

植田喜代志　秋田祥史　荒木雅弘　西本卓也　新美康永

京都工芸繊維大学工芸学部電子情報工学科
〒606-8585 京都市左京区松ヶ崎御所海道町
Email ueda@vox.dj.kit.ac.jp

**あらまし** 本稿では音声対話パターン記述言語であるVoiceXMLをマルチモーダル対話パターンが記述できるように拡張する方法を提案する．この拡張案ではマルチモーダル入出力を2つのフェーズに分けて考える．まず，実際に対話が展開している場面では，マルチモーダル入出力機能は簡易なものに制限することによって，VoiceXMLからの拡張を最小にするように試みている．また，ひとまとまりの対話が終了し，まとまったマルチメディアクリップなどをユーザに提示するような場面では，SMILを用いて豊かな表現ができるようにしている．この方針が本研究室で行なわれているWebコンテンツからの対話パターンの生成に準拠していることを実例を通して示し，拡張案の妥当性を検証した．
**キーワード** VoiceXML,SMIL, マルチモーダル対話システム, 擬人化エージェント

# Towards VoiceXML as a Foundation for Multimodal Dialogue

Kiyoshi UEDA , Masashi AKITA , Masahiro ARAKI , Takuya NISHIMOTO , Yasuhisa NIIMI

Dept. of Electronics and Information Science
Faculty of Engineering and Design, Kyoto Institute of Technology
**Matsugasaki,Sakyo-ku,Kyoto 606-8585,JAPAN**
**Email ueda@vox.dj.kit.ac.jp**

**Abstract** In this paper, we propose an extension of VoiceXML to multi-modal dialogue. The proposed method deals with multi-modal interaction in two stages: in the dialogue phase, the multi-modal extension is limited to simple method so as not to lose control in VoiceXML, and in the presentation phase, SMIL format is used in order to utilize full multi-modal presentation. Now we are examining this approach by implementing whether information multi-modal dialogue according to the guideline of our XML-VoiceXML conversion algorithm.
**Key words** VoiceXML, SMIL, multimodal dialog system, life-like communication agent

## 1 はじめに

近年，コンピュータの進歩に伴い，音声，画像などのマルチメディアデータをリアルタイムで処理できる環境が整ってきた．こうしたマルチメディア処理技術を利用してより自然で効果的な対話を実現できる複数のメディアを利用した入出力を実装したマルチモーダル対話システムが注目されている．また，現在，Web上の情報も多様化しており，これらの情報に対してマルチモーダル対話システムでアクセスできればより便利になる．そこで本研究では音声対話インタフェースを標準化する規格として注目されているVoiceXML(Voice eXtensible Markup Language)[1]をその設計理念を活かしつつ，マルチモーダル対話システムを実現するための拡張案を提案する．この拡張によってVoiceXMLを使えば，マルチモーダル対話システムの専門知識を持っていなくてもマルチモーダル対話システムを構築することが可能になる．

VoiceXMLは，主に電話を利用した音声応答サービスにおける対話パターンを記述するための言語であり，音声対話インタフェースを標準化する規格として注目されている．現在のVoiceXML仕様は電話によるウェブアクセス(音声ポータル)を主な応用範囲として考えられており，画像情報・テキスト情報の提示や擬人化エージェントの制御，マルチモーダル入力には対応していない．元々，人間同士の対

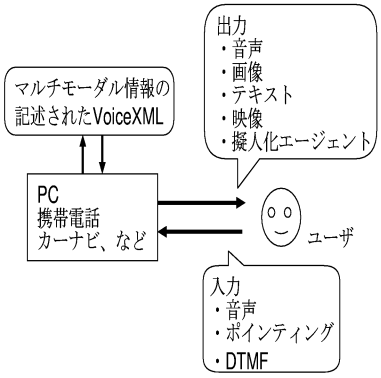


図 1: VoiceXML を用いたマルチモーダル対話システムの概要

話は，音声言語を用いた情報伝達を行うだけでなく，ジェスチャ，視線，表情など，複数のモダリティを適切に使い分けて相手に自分の意思や情報を伝えている．このことから，人間のコミュニケーションは本質的にマルチモーダルであると考えられる．これを，人とコンピュータの対話でも実現できると考えると，複数のモダリティを用いることにより，より自然で効果的な対話が実現できると考えられる．そこで，本研究では VoiceXML に，様々なアプリケーションでマルチモーダル対話システムを実現するための VoiceXML 仕様の拡張について述べる．

最後に本文の構成について述べる．第２章ではマルチモーダル対話の実現方式の検討を行い，３章はマルチモーダル対話システムを実現するために必要な Voice XML の拡張案について，４章では拡張された VoiceXML を用いたマルチモーダル対話システムの応用について記述する．その後，５章で現在の課題・問題点について述べ，６章でまとめとする．

## 2 マルチモーダル対話の実現方式の検討

近年，マルチモーダル対話を記述するため手法として，XML をベースとした SML(Semantic Markup Language)[2] や XISL( Extensible Interaction -Sheet Language)[3] などが提案されている．中でも XISL はコンテンツとインタラクションとビューは独立であるという立場から，コンテンツからインタラクションに関する記述を分離し，独立で利用できるようになっている．しかし，開発者はマルチモーダル対話を実現するためにはコンテンツ・インタラクションの両方を別々に記述する必要がある．

本研究では Web コンテンツはインタラクションを内包するという立場から対話パターン記述言語である VoiceXML にマルチモーダル対話システムを実現するため仕様を加えることにより，マルチモーダル対話システムを実現する手法を取る．本研究では XML から VoiceXML に自動変換することを検討しており，この手法を用いれば，開発者はコンテンツの記述と自動変換された VoiceXML に簡単な手修正を加えるだけでマルチモーダル対話システムを作成することができるという利点がある．

## 3 マルチモーダル対話実現のための VoiceXML 拡張の提案

ここでは，マルチモーダル対話システムを実現するための VoiceXML 拡張の提案を記述する．マルチモーダル対話を VoiceXML を用いて実現するには，音声による入力の他にマウスを用いた入力を許すための機能を加える必要がある．また，出力としてテキストや画像・動画などを表示，擬人化エージェントの制御のための機能も必要となる．本研究ではこれらのテキストや画像などの複雑な表示機能は SMIL(Synchronized Multimedia Integration Language)[4] ファイルを用いて表現し，それらの機能を VoiceXML 処理系の外部の処理系に渡すことによって実現する．しかし，全てを外部の処理系に依存するとインタラクションパターンが消え，対話の制御が難しくなる．そこで，対話のやりとりが成されている間は VoiceXML 処理系で実装できる簡単な入出力を行い，対話が一定の所まで収束した後に，多彩な出力として SMIL ファイルを用いたシステムを提案する．

### 3.1 SMIL を用いた出力

#### 3.1.1 SMIL ファイルの記述方法

SMIL とは Web のストリーミングメディアを統一する言語である．通常 Web で使用されている HTML が，テキストや画像・音声・動画などのリンクを行うように，SMIL はマルチメディア・クリップ統合することができる．SMIL を利用すればプレゼンテーションの時間的な挙動を記述でき，時間軸に沿って，タイミングの指定が可能である．また，スクリーン上のプレゼンテーションのレイアウトを記述することも可能であり，メディアオブジェクトとハイパーリンクを連動できるという特徴がある．図 2 に示すように SMIL ドキュメントは，RealAudio/Video，RealText，RealPix といった個々のマルチメディアクリップを時間的に同期させて再生する様に，それらの表示位置，ファイルの指定，再生方法等を記述する．また，SMIL ファイルの記述方法は XML に準拠したものであり，SMIL ファイルの記述は図 3 のようになる．図 3 の例ではレイアウト情報の<layout>タグ内にそれぞれのクリップの表示場所や背景などの設定を行っている．また<par>タグ内で表示するクリップのファイル名を指定している．これらの記述を変えることにより画像や音声・映像などを様々な形式で出力することがで

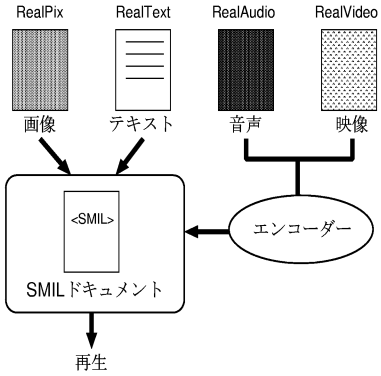


図 2: SMIL の概念図

きるようになっている．

```
<smil>
  <heal>
    <!-- プレゼンテーション情報 -->
      <meta name="title" content="気象情報"/>
      <meta name="author" content="kiyoshi ueda"/>
      <meta name="copyright" content="ueda@vox.dj.kit.ac.jp"/>

    <!-- レイアウト情報 -->
      <layout>
        <root-layout background-color="#bbbbbb" width="500" height="256"/>
        <region id="Image01" left="0" top="0" width="250" height="256" background-color="#000000"/>
        <region id="Text01" left="250" top="0" width="250" height="180" background-color="#000000"/>
        <region id="Text02" left="260" top="200" width="230" height="40" background-color="#000000"/>
      </layout>
  </head>

  <body>
    <!-- クリップソース情報 -->
      <par>
        <ref src="weather.rp" region="Image01"/>
        <textstream src="weather.rt" region="Text01"/>
        <textstream src="weather2.rt" region="Text02"/>
      </par>
  </body>
</smil>
```

図 3: SMIL の記述例

### 3.1.2 SMIL を出力するための拡張

VoiceXML を用いたマルチモーダル対話システムを実現するためのテキスト・画像・動画などを表示の一部には SMIL ファイルを用いる．SMIL は Web のストリーミングメディアを統一する言語であり，テキスト，静止画像，オーディオ，ビデオ，アニメーションを組み合わせたインタラクティブプレゼンテーションが作成できる．SMIL を用いる利点は SMIL の出力を外部処理系に依存することにより VoiceXML 処理系に新たに複雑な処理を加える必要が無いという点にある．画像やテキストを表示する場合，表示時間や表示場所，ウィンドウの大きさなどを決める必要があるが，これらを SMIL ファイルに記述することができる．しかし，出力を全て SMIL に依存してしまうと対話の流れの制御が難しくなる．これは出力を外部に任せてしまうために，VoiceXML の制御を一時的に離れることにより，ユーザの入力を受け付けることができず，システムからユーザへの一方的な情報の流れとなってしまうためである．現在検討している<smil>タグは表 1 のような仕様になっている．図 4 の例は商品 B の説明の SMIL を出力する部分の記述例である．

表 1: smil を出力するための仕様

| 要素名 | 属性名 | 備考 |
|---|---|---|
| smil | src | 必須属性<br>SMIL ファイルの場所 (URI) を指定 |
| | time | 必須属性<br>SMIL ファイルのプレゼンテーションの時間を記述 |

※ obj は商品名を格納する変数とする

```
<field name="ok" type="boolean">
  <prompt>商品 B の説明でよろしいでしょうか
  </prompt>
  <filled>
    <if cond="ok">
      <smil src="b.smi" time="30:00.00"/>
    <else/>
      <clear namelist="obj"/>
    </if>
  </filled>
</field>
```

図 4: smil タグを用いた記述例

## 3.2 マルチモーダル入力を前提とした拡張

### 3.2.1 テキストに関する拡張

テキスト表示は音声対話においてシステムの発話の補助情報として用いる．これは次にユーザの発話できるキーワードを表示したり，システム発話の要約を表示することによって対話をよりスムーズに行うためである．テキスト表示は SMIL を用いた出力でも可能であるが，入力を受け付ける際の補助的な情報として用いる場合には VoiceXML 処理系に実装する方が現実的である．そこで，表 2 に示される仕様の<text>要素を用いることにする．<text>要素は<block>，<if>，<filled>などの子となる．図 5 に示す例では次にユーザが発話すべき情報を提示することによってユーザ発話のための補助情報として使われている．文字の表示開始時間と終了のタイミングについても検討の必要がある．

### 3.2.2 静止画に関する拡張

現在 VoiceXML で実現できる対話システムは，音

表 2: テキスト出力の仕様

| 要素名 | 属性名 | 備考 |
|---|---|---|
| text | size | 文字のサイズを指定<br>指定しない場合はデフォルトの 14 ポイント |
| | color | 文字の色を RGB で指定<br>指定しない場合はデフォルトの黒 |
| | target | target window の id を指定<br>指定しない場合はデフォルトウィンドウ |
| | bgcolor | 文字の背景の色を指定<br>指定しない場合はデフォルトの白 |
| | type | <menu>タグ内でのみ使用可<br>テキストをポインティングで選択可能とする |
| br | | <text>タグ内でのみ使用可<br>テキストを改行する |

```
<field name="pref">
  <prompt>京都府の予報区分には京都府北部と京都府南
  部があります．どちらの情報が必要でしょうか</prompt>
  <text size="14", color="000000">
  京都府北部<br/>
  京都府南部
  </text>
</field>
```

図 5: text タグを用いた記述例

声による応答だけである．これを音声だけでなく，画像を表示しそれをマウスを用いたポインティングで<choice>タグ内の選択ができるようにすると，インタラクションの幅が広がる．また，<field>タグ内でユーザの発話の助けとなるような補助的な情報を表示すれば，よりスムーズな対話を実現できるようになる．例えば，電車の時刻検索のタスクでは，駅名を入力する際に路線図の画像を表示するといったことができる．

画像の表示は SMIL ファイルによって行うことができるが，対話の流れの中で用いるには VoiceXML 処理系にも表示する機能を実装する必要がある．そこで，画像を表示するのは，表 3 に示す<image>タグ及び，属性を用いることとする．

システムの音声出力の補助情報として用いる場合に<block>タグの中で用いたり，<choice >の中で用いれば画像を出力し，マウスでクリックすれば次の対話に飛ぶような入力を受け付けることもできる．図 6 で示す例ではユーザは音声，またはマウスによって alpha.gif もしくは beta.gif の画像をポインティングすることによって対話を進めることができる．

### 3.2.3 動画に関する拡張

表 3: 静止画出力の仕様

| 要素名 | 属性名 | 備考 |
|---|---|---|
| image | src | 必須属性画像<br>画像ファイルの場所 (URI) を指定 |
| | target | target window の id を指定<br>指定しない場合はデフォルトウィンドウ |
| choice | img | 画像を出力し，選択肢として提示することが可能 |
| field | img | 画像を出力し，field に値が埋まれば閉じる |

```
<menu>
  <prompt>A 社のノートパソコンには α タイプと β タイプがございます。どちらになさいますか。</prompt>
  <choice next="alpha.vxml" img="alpha.gif">
  α タイプ</choice>
  <choice next="beta.vxml" img="beta.gif">
  β タイプ</choice>
</menu>
```

図 6: img タグを用いた記述例

動画に関してはほぼ SMIL を用いることになると想定しているが，動画をポインティングの対象とするような場面が考えられる場合は，VoiceXML 処理系への実装を検討する必要がある．現段階での実装案は表 4 に仕様を示す<movie>を考えている．<movie>タグは<block>，<if>，<filled>などの子となる．図 7 の例では地図上を移動する映像 (navi.mpg) による道案内を行い．理解できたかどうかを確認する部分の記述例である．何らかのトラブルで動画が表示出来ない場合は<movie>タグ内のコンテンツが音声で出力される．

表 4: 動画出力の仕様

| 要素名 | 属性名 | 備考 |
|---|---|---|
| movie | src | 必須属性画像<br>動画ファイルの場所 (URI) を指定 |
| | target | target window の id を指定<br>指定しない場合はデフォルトウィンドウ |
| choice | mov | 動画を出力し，選択肢として提示することが可能 |
| field | mov | 動画を出力し，field に値が埋まれば閉じる |

### 3.2.4 擬人化エージェント

本研究では，マルチモーダル対話システムに擬人化エージェントを組み込むことにより，より自然な対話を実現することを目指している．本研究におけ

```
<form id="library">
  <block><movie src="navi.mpg">
  ○○駅北口を降りて東にまっすぐ 300 メートル
  ほど歩いて頂きますと図書館が左手に見えます.
  </movie></block>
  <field name="ok" type="boolean">
    <prompt>ここまではお分かりですか
    </prompt>
    <filled>
      <if cond="ok">
        <goto next="#hischool">
      <else/>
        もう一度説明致します
        <goto next="#library">
      </if>
    </filled>
  </field>
</form>
```

図 7: movie タグを用いた記述例

るエージェントは以下のような機能を持つ.

- ユーザ発話の認識開始を知らせたり，nomatch や noinput イベントで「聞き取れなかった」というジェスチャをしたり，データベース検索中に作業をしている動作するなどのユーザインタラクションの補助
- エージェントが画像を指し示すなどの情報へのハイライト
- システム応答への感情表現の付加

音声合成がエージェントの発話であることを明示するため，speak タグを導入する．また，音声出力に対応するエージェントの動作を play タグで表現する (図 5)．例として play タグを用いて情報をハイライトする方法を説明する．図 8 のように<prompt>要素を記述すると，合成音に合わせてエージェントが画像を指し示す.

現在，エージェントとしては Microsoft Agent を用いる予定である．ただし，表現能力が異なる各種のエージェントの機能を有効に活用し，コンテンツの互換性を損なわないような仕様を目指す.

表 5: 擬人化エージェントの制御の仕様

| 要素名 | 属性名 | 備考 |
|---|---|---|
| play | act | エージェントの動作を記述 |
| | emotion | 動作時のエージェントの感情表現を記述 |
| speak | emotion | 発話時のエージェントの感情表現を記述 |

## 4 マルチモーダル VoiceXML の応用

本研究では，3 章で述べた拡張の実装を行い，実際に拡張された VoiceXML を用いたアプリケーション

```
<menu>
  <prompt>A 社のノートパソコンには<play act="point">
  αタイプ、βタイプ</play>がございます。どちらにな
  さいますか。</prompt>
  <choice next="alpha.vxml" img="alpha.gif">
  αタイプ</choice>
  <choice next="beta.vxml" img="beta.gif">
  βタイプ</choice>
</menu>
```

図 8: play タグを用いた記述例

の実装を行っている．現在はマルチモーダル情報の記述された XML ファイル (図 9) を VoiceXML ファイルに変換しマルチモーダル対話を実現している．マルチモーダル情報は，コンテンツタグに囲まれた画像や SMIL ドキュメントをそのまま表示，または再生する場合は XML のリンク先にあるファイルをそのまま用いて VoiceXML の中に記述する.

```
<?xml version="1.0" encoding="Shift_JIS"?>
<天気情報 地域="近畿">
  <地域 都道府県="大阪" 日付="明日">
    <天気>晴れ</天気>
    <最高気温>26 ℃</最高気温>
    <最低気温>20 ℃</最低気温>
    <降水確率 時間帯="午前">10%</降水確率>
    <降水確率 時間帯="午後">0%</降水確率>
  </地域>
  <地域 都道府県="京都" 日付="明日">
    <天気>晴れのち曇</天気>
    <最高気温>25 ℃</最高気温>
    <最低気温>20 ℃</最低気温>
    <降水確率 時間帯="午前">20%</降水確率>
    <降水確率 時間帯="午後">10%</降水確率>
  </地域>
          .
          .
          .
  <天気予報地図 >kinki.gif</天気予報地図>
  <アメダス>kinki2.gif</アメダス>
  <Smil src="./kinki.smi">天気ニュース</Smil>
</天気情報>
```

図 9: XML で記述された気象情報

現在，マルチモーダル対話システムを気象情報のタスクを用いて，図 10 に示すような VoiceXML アプリケーションの実装を行っている．XML で記述された Web ページが少ないので，実際の Web 上にある HTML ファイルを，まず XML ファイルへ変換して使用している．この XML ファイルに記述されたマルチモーダル情報を，Java サーブレットを用いて動的に VoiceXML ファイルに変換しユーザの要求に答える出力を生成するものである.

このツールを用いて動的に生成した VoiceXML ファイルを図 11 に示す．現在の所，VoiceXML 処理系で実装されている<img>タグのみを使用した簡単なものであり，最初に 5 行目の weather1.gif で天気図の画像を表示しながら天気の情報をシステムが発話し，降水確率を伝える直前で 8 行目の weather2.gif でアメダスの画像に切替えて表示し，この VoiceXML

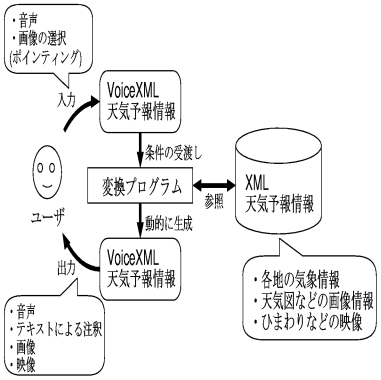


図 10: 気象情報検索システムの概要

ファイルの処理が終了すると同時に画像が閉じられるようになっている．ファイルの最後の部分で更に詳しい情報が必要かどうかをユーザに質問し，必要であれば SMIL ファイルを用いた更に詳しい出力を用いることを予定している．このように XML から VoiceXML へ自動変換を行うと，開発者はコンテンツのみを記述し，自動生成された VoiceXML ファイルに簡単な手修正を加えることによってマルチモーダル対話システムを作成することが可能となる．

```
<?xml version="1.0" encoding="Shift_JIS" ?>
<vxml version="1.0">
  <form id="お天気情報">
    <block>おまたせしました</block>
    <block><img src="kinki.gif" target="1">
    </block>
    <block>大阪府の明日の気象情報をお送りしま
    す</block>
    <block>大阪府の明日の天気は晴れ，最高気温
    は 26 度最低気温は 20 度です</block>
    <block><img src="kinki2.gif" target="1">
    </block>
    <block>降水確率は午前が 10%，午後が 0%です.
    </block>
    <field name="temp" type="boolean">
      <prompt>
        更に詳しい情報が必要ですか，はい，か，
        いいえ，で答えてください.
      </prompt>
      <filled>
        <if cond="temp">
          <goto next="./weater_2.vxml">
        <else/>
          ご利用ありがとうございました.
        </if>
      </filled>
    </field>
  </form>
</vxml>
```

図 11: マルチモーダル対話を実現する VoiceXML

## 4 問題点の検討

本研究ではマルチモーダル対話を実現する Voice-XML の拡張案を提案したが，現時点ではまだ，様々な問題点が有る．まず，テキスト・画像・映像については表示のタイミングや表示時間，ウィンドウの大きさなどの指定については，まだまだ検討が必要になる．また，画像を 1 つのオブジェクトとして扱っているので，画像中の要素に対するポインティング入力が扱えない．しかし，仕様が複雑になれば，タグの種類なども増え，開発者の新たな負担となってしまう．逆に，単純な仕様にするとインタラクションのパターンが限られることになり，多彩なマルチモーダル対話を記述することができなくなる．また，SMIL ファイルは多彩な出力を行えるが，現在の仕様では VoiceXML の処理と離れた所で動作するために出力のみにしか対応しておらず，マルチモーダル入力を行うことができない．今後は SMIL を用いたマルチモーダル入力 [5] への対応も検討する．

## 5 おわりに

本研究では，マルチモーダル対話システムを実現するための VoiceXML の仕様の拡張についての提案を行い，その問題点についての検討を行った．また，実際に気象情報のタスクを用いたマルチモーダル対話システムの実装も進めつつ，マルチモーダル対話システムに必要な機能の選定を行っている．現在は<img>タグと<text>タグの一部のみが実装されているが，今後はその他のタグに対する実装も順次行っている．同時に，複数のタスクのマルチモーダル対話システムの実装も行いつつ，マルチモーダル対話システムを実現するための VoiceXML の仕様の拡張案を検討し，よりスムーズに対話のできる仕様を提案することが目標となる．

## 参考文献

[1] VoiceXML1.0(2000): http://www.w3.org/TR/voicexml/

[2] Kuansan Wang: Implementation Of A Multimodal Dialog System Using Extended Markup Languages. Proc. ICSLP2000; 1150-1156. (2000)

[3] 小林 聡, 中村 有作, 桂田 浩一, 山田 博文, 新田 恒雄: マルチモーダル対話記述言語 XISL の提案. 音声言語情報処理 37-8; 43-48. (2001)

[4] SMIL2.0(2001): http://www.w3.org/TR/smil20/

[5] Jennifer L. Beckham, Giuseppe Di Fabbrizio, Nils Klarlund: Toward SMIL as a Foundation for Multimodal, Multimedia Application. Eurospeech2001; 1363-1367