Web ブラウザをインターフェースとした マルチモーダル対話システム

新田恒雄，桂田浩一（第 7 工学系，e-learning コア）
入部百合絵（情報メディア基盤センター，e-learning コア）

1. はじめに

人とコンピュータ間のマルチモーダル対話（Multi-Modal Interaction: MMI）を実現するために，数多くのシステムが開発されてきた [1) 2)]．これまでに提案されたシステムは，特別なソフトウェアのインストールを必要とする，あるいはジェスチャ・表情解析のための特別な装備を要請するなど，導入コストがネックとなり，一般ユーザにまでは MMI が浸透していない現状がある．

我々は MMI の持つ利便性を広く普及させることを目標に，現在，広く使われている Web ブラウザをインターフェースとした MMI システムを開発した．本システムは，JavaScript などの標準技術のみを用いており，特別なソフトウェアのインストールや高性能端末の装備を必要とすることなく MMI を実現している．

以下，2. で本システムのベースとなる Galatea for Windows [3)]について説明した後，3. 以降では本システムの構成を述べる．

2. Galatea for Windows の構成

Galatea for Windows は Galatea Project により公開されている MMI システムであり，音声やエージェントを用いた自然な対話をコンピュータとの間で行なうことを可能にしている．表 1 に Galatea for Windows が扱えるモダリティを，また図 1 にモジュール構成を示す．

図中の対話制御部は，XISL [4)]で記述された MMI の対話シナリオを解釈・実行するモジュールである．XISL を含む XML 文書は，ドキュメントサーバ上に置かれている．

フロントエンドはユーザとのインターフェースを担うモジュールで，①ウェブページを表示するブラウザ，②音声認識エンジン Julius，③顔画像合成エンジン GalateaFSM，④音声合成エンジン GalateaTalk から構成されている．

フロントエンドと対話制御部の二つのモジュールを連携させることで，ユーザがエージェントと対話しつつ，ブラウザ操作やウェブページ閲覧を行なうことを可能にしている．

表1 Galatea for Windows で扱えるモダリティ

	モダリティの種類
入力	音声，ポインティング，キーボード
出力	ブラウザ，エージェント，合成音声

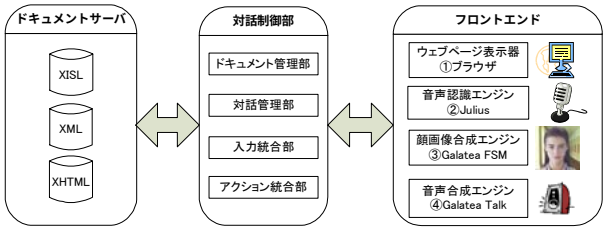

図 1 Galatea for Windows のモジュール構成

3. システムの構成

音声認識や顔画像合成をブラウザ単体で低負荷に実装するのは困難である．そこで我々は，Ajax および Comet の技術を利用して，サーバとウェブブラウザを連携させ，高負荷な処理をサーバ上で，低負荷な処理をブラウザで行なうようシステムを設計した．図 2 にシステムの構成を示す．以下では，ウェブブラウザでの処理とサーバ上での処理に分けて説明する．

3.1 Web ブラウザでの処理

3.1.1 ユーザからの入力取得

音声認識は複雑な処理を必要とするため，ブラウザ単体で低負荷に実装するのは困難である．そこで本システムでは，録音をウェブブラウザで，認識をサーバ上で行なうことにより，ブラウザへの負荷が少ない音声認識を実現した．この手法は西村らの w3voice [5)]で実用性が確認されている．録音は Java Applet を用いた音声録音器 Sound Recorder で行なった．音声は Base64 エンコードの後，サーバに送られる．ポインティングなど音声以外の入力は JavaScript を用いて取得している．

3.1.2 ユーザへの出力

顔画像合成と音声合成によるエージェント出力は高負荷な処理となる．そこで本システムでは，顔画像と音声をサーバ上で一つの動画に結合し，ブラウザの動画再生機能でエージェント出力を低負荷に実現する方法を採った．動画再生には Adobe Flash を用いて実装した Agent Presenter を使用した．ページ遷移などブラウザへの出力は JavaScript によって実現している．

3.2 サーバ上の処理

ウェブブラウザからのマルチモーダル入力データは，図 2 に示すサーバ上の Session Manager が受け取り，統合前処理（音声入力は音声認識結果に変換）の後，Dialog Manager 内の Input Integrator で統合処理が行なわれる．統合結果は XISL Interpreter で解釈され，対話シナリオに沿って出力命令が生成される．出力がエージェントの場合は，Agent Manager で動画が生成され，Session Manager を通じてウェブブラウザに送信される．

その他の出力（ウェブページの表示など）の場合，そのまま Session Manager を通じてウェブブラウザにウェブページ表示などの命令が送られる．

4. コンテンツの作成

本システムを用いて MMI コンテンツを作成するには，(i) 対話シナリオを記述した XISL 文書を作成し，(ii) ウェブページを記述した XHTML 文書に幾つかの JavaScript を組み込めば良い．JavaScript によって，ウェブページ内に自動的に Sound Recorder と Agent Presenter が埋め込まれる．図 3 にショッピングサイトの開発例を示す．このウェブページの例では，エージェントが商品説明を行い，音声入力で商品の選択や購入が可能になっている．

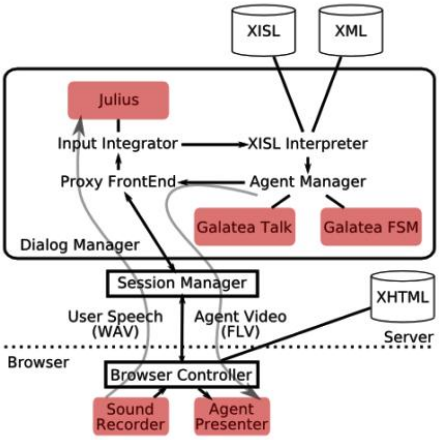

図3 MMI システムの構成

図2 MMI システムを組み込んだショッピングサイト

5. おわりに

本稿ではウェブブラウザをインターフェースとした MMI システムについて述べた．本システムではブラウザとサーバを連携させることにより，低機能な端末でも実行可能な MMI システムを実現している．今後は携帯電話や PDA などの携帯端末でのシステム動作の検証やクロスブラウザ対応を目指していきたい．

参 考 文 献

1) Seneff,S., et al,. “Galaxy II: A Reference Architecture for Conversational System Development”, Proc. Of ICSLP’98, pp.931-934 (1998).
2) N. Reithinger, et al., “SmartKom adaptive and Flexible Multimodal Access to Multiple Applications”, Proc. Of ICMI’03, pp.101-108 (2003).
3) S. Kawamoto, et al., “Galatea: Open source software for developing anthropomorphic spoken dialog agents”, in Life-Like Characters, ed. H. Prendinger and M. Ishizuka, pp.187-212, Springer-Verlag (2004).
4) 桂田,他,”MMI 記述言語 XISL の提案”,情報処理学会論文誌,Vol.44,No.11,pp.2681-2689 (2003).
5) 西村竜一，他，”音声入力・認識機能を有する Web システム w3voice の開発と運用”，情報処理学会研究報告，2007-SLP-68-3，pp.13-18 (2007).

発 表 論 文

（1） Mohammad Huda, Hiroaki Kawashima, Kouichi Katsurada and Tsuneo Nitta: "Distinctive Phonetic Feature (DPF) Based Phoneme Recognition Using MLNs and Inhibition/Enhancement Network for Noise Robust ASR", Proc. of NCSP09, 1PM2-K3 (2009-3).
（2） Kouichi Katsurada, Teruki Kirihata, Masashi Kudo, Junki Takada and Tsuneo Nitta: "A Browser-based Multimodal Interaction System", Proc. of ICMI'08, pp.195-196 (2008-10).
（3） Mohammad Nurul Huda, Kouichi Katsurada and Tsuneo Nitta: "Phoneme Recognition Based on Hybrid Neural Networks with Inhibition/Enhancement of Distinctive Phonetic Feature (DPF) Trajectories", Proc. of INTERSPEECH2008, pp.1529-1532 (2008-9).