詳細説明資料

音声対話からマルチモーダル対話へ

■　音声対話システムは，電話回線を通して利用者がコンピュータと音声で対話する装置として，1970年代に始まりました。多くの方が，電話のプッシュボタンと音声応答で，チケット予約の問い合わせなどをされた経験をお持ちでしょう。近年は，電話とコンピュータが一体となって，自動で案内業務を行う，ボイスポータルサイトも増えています。

音声対話では，音声認識結果(「豊橋から東京まで切符を三枚ください」)から発話内容を解釈します。この例では，出発地=豊橋，目的地=東京，切符枚数=3，大人/子供=? となります。次に解釈結果に応じて，利用者に返す応答を作成します。例では，大人か子供かを確定しないと切符を発行できません。そこで，システムは「大人の枚数と子供の枚数を指定してください」というメッセージを返します。このように，入力音声の内容に応じて応答文を作成するといった，一連の流れをシステムの中で予め用意する必要があります。一連の対話の流れを記述する言語として，現在では VoiceXML という音声対話記述言語が，インターネットの標準化団体 W3C で標準化されています。

■　一方，インターネットの世界では，マウスクリック(あるいはキーボードやタッチ操作)と表示画面の組み合せで対話が構成されてきました。これをグラフィカル・ユーザー・インタフェース(GUI)と呼びます。携帯情報端末の時代に入り，こうした GUI と音声対話を合体させる必要が出てきた結果，マルチモーダル対話(略称 MMI; Multi-Modal Interaction)が脚光を浴びています。

様々な入力と出力を同時に扱うことで，MMI は次のような多くの利点を持っています。

(1) 分かり易い対話を提供できる:　例えば，「画面表示とタッチ操作」および「音声によるガイドと音声入力」の双方が提供されていれば，利用者は自分の目的に合った組合せを使用できる。

(2) 対話を円滑に進行できる:　例えば，「グラフを表示しつつ(一覧できる)，音声でグラフの詳細を説明する」，あるいは，音声入力とタッチ入力を並行して利用することで，周囲が煩い時はタッチで，操作項目が多くて目では追えない時は音声入力で，というように環境の変化に応じた利用を可能にします。

(3) 人間に近付いた対話を提供できる:　「これ」と言いながら指で図を指し示す，といった人間が使用することの多い対話方法を利用できる。「画面にアニメのキャラクタを登場させ，このキャラクタ(エージェントと呼ばれます)との間で，互いに音声やジェスチャを使用して対話する」ことで，人間に近い対話を行える。

このほか表 A に示すように，音声入力とペン(あるいは指)入力などは互いに相補う機能があり，こうした複数の入力操作を適宜利用できるようにしたことも，MMI の利点と言えます。

表 A　ペン入力と音声入力
~ 二つの入力操作の相補性

項　目	ペン入力	音声入力
利用者の拘束	•パッドの上に構えて操作(目と手を拘束する)	•動作は拘束しない
	•利用場所の制限は少ない	•会議中，騒音下で使えない
入力速度	•遅い	•速い(健常者の場合)
記録・編集	•記録に残り，編集も簡単	•記録・編集には不向き
入力対象	•少項目の確実な直示に向く	•多項目の直示が可能
	•文字・図形・ジェスチャと多彩な機能を持つ	•感情・個人性を表現し易い
その他	•考えながらの入力に適する	•即応的な使い方に適する
	•聴覚・発声障害者も利用可	•視覚障害者も利用可

マルチモーダル対話を制御する言語 XISL 開発の経緯

■　マルチモーダル対話(MMI)は，上に述べたように音声だけの対話と比べ，より複雑な対話制御が必要になります。この研究は，新田が(株)東芝で’80 年代に様々な音声入力システムを開発した際に限界を感じたことから始まりました。当時開発したシステムには，TV の音量・チャンネル切り替えのための音声認識ボード開発，銀行向け電話音声認識・応答システム開発，音声認識 LSI と音声入力電話機の開発，様々な社会システム(エレベータシステム，券売機，銀行 ATM ほか)への音声入力応用などが挙げられます。

■　‘90 年代に入って，音声入出力だけの限界を確信し，画面とタッチ操作を加えたシステム(マルチモーダル対話システム)の研究に重心を移すと共に，複雑になる対話制御を行う言語の研究を開始しました。この頃開発したシステムには，警視庁の依頼で開発した地理案内システムなどがあります。利用者の接近を感知するセンサー，利用者のタッチ操作と自由発話(「えーと，ヒルトンホテルは何処ですか?」など)を音声で入力できるシステムで，案内地図は FAX で出力しました。

■　‘98 年末には豊橋技術科学大学へ移り，新たにマルチーダル対話システムの開発と，開発に必要な MMI 記述言語の研究を再開しました。この頃，インターネットの標準化団体である W3C (World-Wide-Web Consortium) では，XML (eXtensible Markup Language) を音声対話の制御に使用する，音声ブラウザ(Voice Browser)のワーキンググループが発足しています。そこで，マルチモーダル対話(MMI)の記述言語が，標準化日程に上ること見越し，XML ベースの MMI 記述言語を開発することにしました。

■　2001 年から 5 年間は，W3C の MMI-WG へ参加し，マルチモーダル対話の標準化活動に従事しました。この間，研究室では XISL (eXtensible Interaction Scenario Language) の開発と実装・改良に努めました。この頃はマイクロソフト社，インテルなどが押す言語 SALT や，IBM などが押す X+V が互いに推進グループを形成して競いましたが，現在，これらの活動は両者とも下火になっています。

XISL が他の言語に比べ優位な特長は，後述するように，(A) 最初からマルチモーダル対話制御に必要な記述能力を持たせていたこと，および(B) 入出力(モダリティと呼ばれる)の拡張性が高く，新しい端末仕様にも対応できること，の二つが挙げられます。

■　2003 年から 2007 年にかけて，情報処理学会の音声対話技術コンソーシアム(ISTC: Interactive Speech Technology Consortium) 活動を立ち上げ，代表として音声・マルチモーダル対話研究の普及活動に従事するともに，同じく情報処理学会の情報規格調査会学会で試行標準化専門委員会の活動を通して，音声インタフェース，マルチモーダル対話に関する標準化を推進しました。

図 C は，コンソーシアム活動の中で開発したマルチモーダル対話システムの開発ツール(Interaction Builder) です。開発に必要な入出力の部品を登録しておくことで，マウス中心の簡単な操作により複雑なシステム開発ができるようになりました。

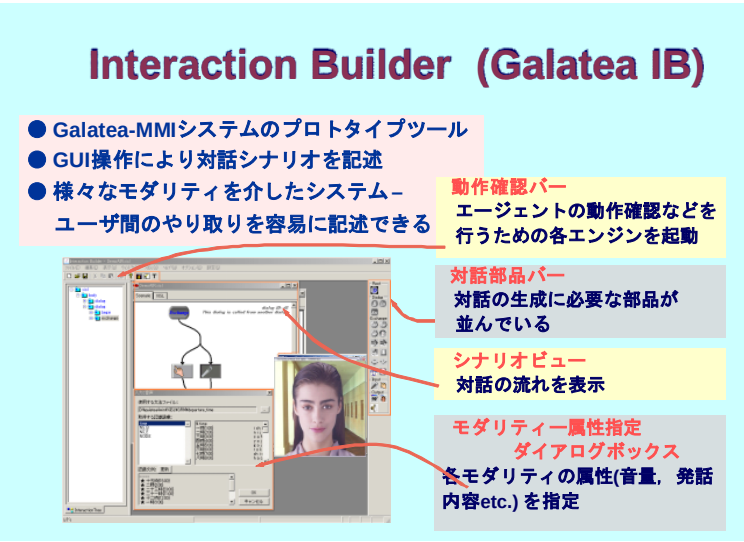

図 C: マルチモーダル対話システム開発ツール

マルチモーダル対話の制御言語 XISL の特長

■　マルチモーダル対話に必要な対話制御を記述する能力に優れています

マルチモーダル対話では，様々な入出力を適切に制御することが必要です。一例として，図 D に三種類の制御を示しています。

(1) 逐次的制御　これは入出力を順番に制御する要素です。最初に商品を入力し，続いて個数を入力するなどの場合です。全て入力しないと次には進めません。
(2) 並行的制御　これは入出力を全て並行して(並列に)制御する要素です。順番は任意です。商品と個数を順不同で受け付けられます｛「りんご(を)」→「三個」,「三個(の)」→「リンゴ」など｝。逐次的な処理と比べて効率良い対話ができますが，入力要素が足りない場合は，催促などの処理が必要になります。
(3) 択一的制御　これは入出力を択一的に制御する要素です。すなわちどれか一つが受付けられると，次に進みます。音声入力でもタッチ入力でも，またその混合でも同じ「りんご」を指示しているなら，受け付けます。

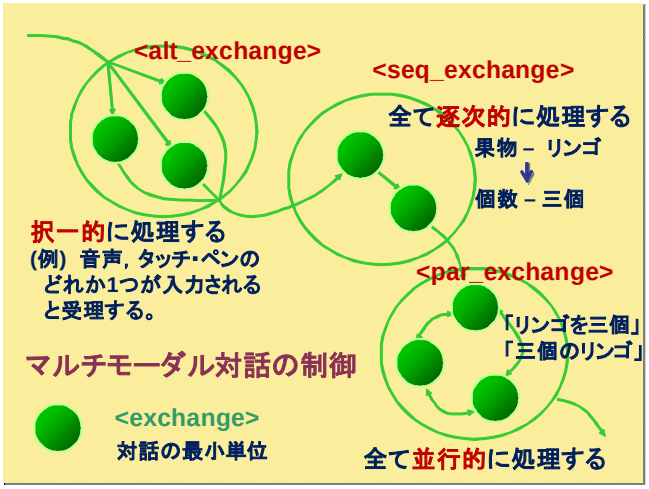

図 D: マルチモーダル対話における
三つの制御　｛逐次，並行，択一｝

XISL は，条件分岐や入れ子構造によって，さらに複雑な対話を記述することも可能になっています。

■　入出力(モダリティと呼ばれる)の拡張性が高く，新しい端末仕様にも対応できます

この特長を実現するために，入出力を記述する要素(<input> および <output>) は，内容の詳細を端末や入出力方法(モダリティ)ごとに自由に規定できるようにしました。
端末の入出力インタフェース(フロントエンド)は，対話を開始すると対話制御部(XISL が制御)との間で，例えば GPS 情報では，属性値が「GPS」，受け取る変数は「位置情報」のように動作を規定して対話を進めることができます。

これまで，XISL は様々な応用システム上で実装・評価が行われてきました。今後の計画としては，以下があります。
(1) 情報処理学会試行標準化委員会が策定した「マルチモーダル対話の 6 階層モデル」に基づく，対話システムの実装と評価（デモ予定；より柔軟な対話システムの開発が可能に)。
(2) 様々なセンサーネットワークからの動的情報を端末で受け，状況に則した対話行動を支援できるシステムの設計と実証テスト(カーナビやロボットなどの適応行動が可能に)。