音声言語情報処理 7−11
(1995. 7. 21)

マルチモーダルインタラクションシステムの試作

宮崎敏彦　須崎昌彦　久野裕次　田川忠道
miya@kansai.oki.co.jp
沖電気工業(株) 研究開発本部

人とコンピュータの円滑なインタフェース構築を目指して、コンピュータグラフィックスによる顔画像生成と、画像認識, 音声対話を統合したシステムを試作した。画像認識部ではディスプレイの前の人物の頭部位置を検出し、検出結果を入力画像毎に対応付けることによって顔画像の視線の動きを自然なものにしている。音声対話部では対話の状態にあった入力手段を提供するなど、ユーザが適切な応答をすることができるサポートをすると同時に、顔画像の表情や動きを変化させることで対話システムの欠点を補っている。さらに、システムの機能を補間するという位置付けで、デスクトップTV会議システムと結合し、システムが対処できない状況では適宜専門化に補助を依頼することもできる。

A Prototype Multi-modal Interaction System

Toshihiko Miyazaki　Masahiko Suzaki　Yuji Kuno　Tadamichi Tagawa
Research and Development Group, Oki Electric Industry Co.,Ltd.

For the purpose of easing human computer interaction, we built a visitor guidance system integrating facial animation by computer graphics, image processing, and speech dialogue. Gaze directions of the facial animation are controlled by detecting head positions of the persons in front of the display and tracking the person who is regarded as the main target. The speech dialogue part gives a user appropriate answers by incorporating an extended plan reasoning method. By changing facial expressions and movements of the facial animation, we can show the states of the system to a user as nonverbal information to make up for the weakness of the speech dialogue such as inaccuracy of voice recognition. We integrated the desktop conference system into our multi-modal interaction system. It can decrease unsolved situations with the assistance of a human expert.

1 はじめに

人は主に言葉を使ってコミュニケーションしているが、電話や電子メイルといった言葉だけのコミュニケーションでは、互いの意図が伝わりにくいというもどかしさを感じることがよくある。これは互いに向かい合った対話では、言葉以外のメディア、例えば絵や身振りあるいは顔の表情や目の動きなど、いわゆるノンバーバルな情報伝達手段を、我々がうまく併用できているためであろう。このような見地に立って、近年HCIの分野ではマルチモーダルインタフェースの研究が盛んに行なわれるようになってきた[1][7][12]。

マルチモーダルインタフェースの設計方法としては、人と人の間で行なわれている実際の対話を観察することによって必要な機能等を洗い出す方法と、試作と試用を繰り返すことによってより洗練されたインタフェースを確立していく方法が考えられる[1]。一方、種々のモードを実現するための、例えば音声認識など、個々の技術はまだ完全なものとは言えず、他のモードとの統合によって機能の補完を考える必要がある要素技術も多い。

従って我々は、それぞれの要素技術がどの程度何に有用か、あるいはモードの組合せによってどのような機能補完が可能かなど、研究のアプローチとしては後者の進め方が重要ではないかと考え、マルチモーダルインタラクションシステムのプロトタイプを試作している。具体的には、音声対話とCGによる顔アニメーション, 画像認識等を統合した来訪者案内システムの構築と評価を現在進めているところである。

本論文では、我々が構築中の来訪者案内システムの概要と、試作を通して得られた知見に基づく今後の課題について述べる。

2 マルチモーダルインタラクション

マルチモーダルインタフェースの研究の目的は、人とシステムとのインタラクションを円滑に進めることのできる手段を提供することである。そのためには、複数のモードをいかにバランス良く統合するかが重要となる。言い換えると、どのメディアにどのような機能を持たせるか、あるいは何と何を複合させて機能を提供するかといったことが十分洗練されていなければ、実際に使いやすいインタフェースとはならないだろう。しかし一方では、例えば音声認識など、新しいモードを実現する上で重要であるような要素技術が、まだ十分成熟していないという問題もある。このためシステムを設計するためには、どのような要素技術がどの程度使えるか、あるいはまた、要素技術の統合によってどの程度機能が向上するのかを確かめながら進めていく必要があると考えられる。

以上のような見地から、我々は来訪者案内システムと呼ぶマルチモーダルインタラクションシステムを実際に試作し、これに種々のモードを付加し評価しつつより良いインタフェースを模索するというボトムアップ的なアプローチで、マルチモーダルインタフェースの研究を進めている。

プロトタイプとして設定した来訪者案内システムの、インタフェースに関する基本方針は次の点である。

- 来訪者案内という状況と、人と人の自然なコミュニケーションの手段が声であることから、音声によるコミュニケーションを中心とし、これを補完するために他のモードを統合する
- 音声による計算機とのコミュニケーションには抵抗感を持つ人もいる。対話の自然さとモチベーションの向上のために、計算機には表情を持った顔を持たせる。
- 対話者の当事者意識を高めるために、システムには外部を観察する目を持たせ、対話者の方向などの視覚的な認識機能を持たせる。(外部を観察する機能により、例えば[11]や[12]に示されているようなより高度なサービスを提供できる可能性もある。)
- 来訪者案内という状況を限定しても、あらゆる状況に対応できるシステムを構築することは不可能である。そこで、システムの不完全さを補うために、困った状況では人間が代わって対処できるような機構を組み込む。具体的には、我々が別途開発中のデスクトップTV会議システム[14]を結合することによってこの機能を実現する。勿論、来訪者案内本来の目的にもこの機能は有用である。

3 来訪者案内システム

3.1 システム概要

先にも述べたように、我々は音声対話システムと顔アニメーションに、画像認識部を加え、オフィスの受け付け等で来訪者の案内を行なう来訪者案内システムを開発している。図1にシステムの構成を示す。テレビカメラはディスプレイ上部に置かれ、画像処理によりディスプレイの前にいる人物の頭部位置を検出し、システム側から見た来訪者の位置や動きの情報として顔

[1]勿論実際には両者を繰り返すことになろうが、ここでは主なアプローチを述べている。

アニメーションの視線方向などを決定する。また、音声対話処理によって来訪者の用件を聞きだし、用件を満たすための適切なサービスを来訪者に提供する。また、対話の状況によって顔アニメーションの動きや表情を変化させ、システムの状態をノンバーバルな形で来訪者に伝える。

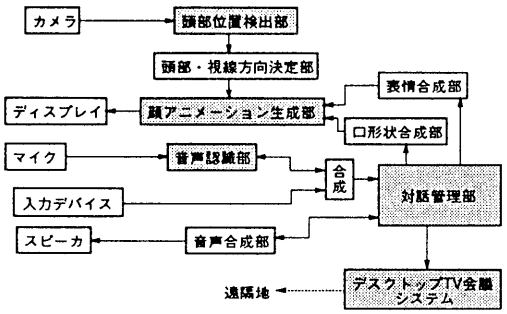

図 1: システムブロック図

来訪者が誰かに面会を希望している場合や、システムだけでは対処しきれず人の助けが必要であると判断された場合は、デスクトップTV会議システムを使って適当な人間を呼び出し、案内システムは補佐役に回る。

図2は、来訪者案内システムの外観である。

以降では、図1にマスクの掛かった各部に関し簡単に説明を加える。

3.2 対話管理部

対話管理部は、各サブシステムを、対話の状況に応じて制御する部分である。

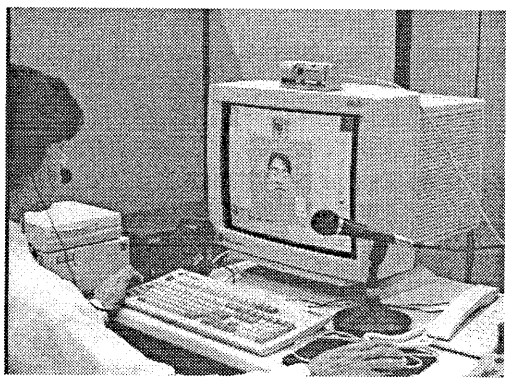

図 2: 来訪者案内システムの外観

来訪者案内システムでは、音声によるコミュニケーションが情報伝達の中心手段である。しかし実際には、地図やメニューの例などでも分かるように、視覚など他の情報伝達手段を使った方が良い場合もある。従って対話管理には、その進行状況に応じて適切と考えられるコミニュケーション手段を適宜選択できる枠組が必要である。また音声対話を実現する上では、発話中の省略表現や指示代名詞の照応といった自然言語が持つ問題も解決するメカニズムが必要である。

このため我々は、対話管理方法として[4]等で提案されているプラン推論を用いた方法を採用し、これに以下のような拡張を行なった。

1. コンテキストに応じた入力手段の選択機能

2. 応答候補の遅延提示機能

3.2.1 入力手段の選択

対話処理の基本的な流れは、人の発話を構文解析、意味解析およびプラン認識を経て発話意図を抽出し、次に、この意図とあらかじめシステムに与えられている対話のプランスキーマをもとに、人の発話意図を満足させるためのシステム側の応答を生成する、というサイクルを繰り返す。

例えば、来訪者が「××さんお願いします」と言ったとしよう。プランスキーマとして、『取り次ぐ場合は来訪者の名前を知る必要がある』と定義されていると、相手の名前を尋ねる次のような応答候補(ゴール)が得られる。

ask_value(system, user, 名前)

この場合、当然ながらシステムは相手の名前を知らない。すなわち、人が音声で答えるとしたとき、人の発話内容を予測することができない。一方、現状の音声認識技術では、予測できない文字列を認識することは非常に困難である。従って、この場合はペンかキーボードを使って答えてもらうのが適当であろう。

このように、会話のコンテキストに応じて、適切な入力手段が提供できるようにするために、システムの次発話を表すゴールとその引数の型を元に、ユーザに提供すべき入力手段を決定する機構が組み込まれている。上の例の場合では、プランスキーマの記述者は、"名前"という属性の値が、要素の限定できない集合を持つ型であると宣言しておけば良い。システム側では、ゴールの述語と型のペアに対し、どのような入力手段を提供すべきかという組が定義されており、これをもとに入力手段が決定される。

3.2.2 応答候補の遅延提示機能

システムが来訪者に対し、どの部署に用があるかを問い掛ける場合のように、システムにとって質問の多くは選択的なものである。このとき、どのような問い掛け方をするかは、ユーザインタフェースの観点から重要である。常にメニューを提示して選択してもらうのでは、画面が繁雑になるとか、ルーチン作業になり熟練者が面倒に感じるといった問題があろうし、何も提示しないと、名前を断片的にしかしらないといった場合に答えられないという問題がある。

これに対し我々のシステムでは、人からの自発的な音声応答を優先し、人が答えられない場合に、メニュー提示によってそこからの選択を促すという方式を採用している。試作システムでは、次の場合に人が答えられないと判断している。

- 問い掛けに対し一定時間応答が無い
- 音声認識に2回以上失敗した

また提示するメニューは、プランスタック中で予測されているゴールを元に、これを満足する値の集合を実行時に決定することによって生成している。

3.3 顔アニメーション生成部

顔アニメーションのための顔の基本モデルは、3次元のワイヤーフレームモデルに、人の顔写真をテクスチャとして張り付けることによって作成する。この3次元モデルを、以下で述べるような筋肉モデルを用いて適宜変形することによって様々な表情や発話時の口の動きを合成する。

人の表情は、顔面の筋肉、すなわち表情筋が収縮することによって表出する。Watersはこの表情筋をモデル化し、計算機上で容易に筋肉の動きをシミュレートできる筋肉モデルを開発した[3]。筋肉モデルには、線状筋、括約筋、方形筋の3つのタイプがあり、これらを顔の3次元モデル上に適切に配置することで表情を作成する。

図3は、Watersの線状筋肉モデルを拡張したモデルを示している。この筋肉が収縮することによって、図中の$V_1P_rP_r'$で囲まれる領域内(以下、筋肉の影響範囲と呼ぶ)の頂点が筋肉の始点V_1の方向へ引き寄せられる。Watersの線状筋肉モデルでは筋肉の影響範囲が、直線V_1V_2に対して対称であるのに対し、拡張モデルでは直線V_1V_2に対して影響範囲を非対称に設定できることが特徴である。これにより、筋肉の収縮による影響範囲に関し、さらに細かな設定を行なうことができる。

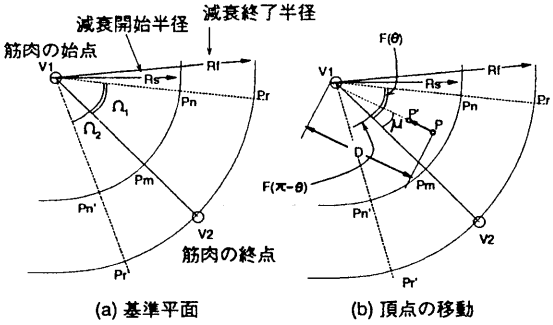

図 3: 線状筋肉の拡張

基本的な感情を表出する基本表情(喜び、怒り、悲しみなど)については、EkmanらのFACS(Facial Action Coding System)[2]を参考にした。さらに、問い掛けや思案など、対話の過程で表出される表情を作成した。(詳しくは[13]を参照されたし。)

3.4 頭部位置検出部

ディスプレイの上に設置されたカメラは、コンピュータの目に相当する。顔画像がユーザの頭部位置によって視線方向を変化するように、入力画像中の人物の頭部位置を検出する。

入力画像にはユーザ以外に複数の人物が存在する場合や、ユーザが2人以上の場合が考えられるので、頭部位置の検出は、

1. それぞれの人物に対応した人物領域の切り分け
2. 各人物の頭部位置の検出
3. それぞれの人物固有のラベル付け

という3つの処理を順次行なう。

また、画像処理による人物の頭部位置検出の方法としては色情報[9]や濃淡情報[8]を用いたパターンマッチングによる方法が報告されているが、実時間での処理には専用のハードウェアを必要とする。従ってここでは、入力画像と背景画像の差分画像から得られる、人のシルエット像の形状から頭部の検出を行なう方法を用いる。(詳細は[13]参照。)

図4は人物領域の切り分け、頭部位置検出、ラベル付けを行なった結果である。二重の□印がラベル1の人物の頭部位置、□印がラベル2の人物の頭部位置を示している。ラベル1とラベル2の人物が重なる前後のフレームでもラベル付けが正しく行なわれている。

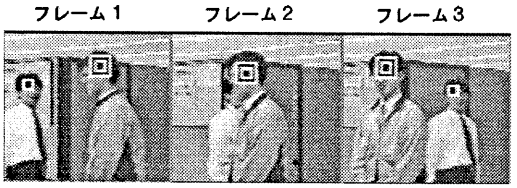

図 4: 頭部位置検出処理結果

3.5 音声認識部

音声認識部は、ソフトウェアのみで実現されたHMMにもとづく不特定話者連続音声認識プログラムである。認識速度は、発話時間や認識時に用いる文法の規模によって異なるが、1秒から3秒程度の時間で認識できる。しかし実際の対話では、3秒もの間システムから何の応答もないと、人が不安感や不快感を持つため、認識部は認識に必要なおおよその時間を即座に対話管理部に送る。対話管理部では、この長さに応じて、あいづちやうなずきなど適当な応対を、顔アニメーション生成部などに促す。

また、マウスやタッチパネルなど音声以外の入力モードとの複合で情報の入力ができるよう、認識結果の個々の単語に対し、その発話開始時刻と終了時刻が付加されており、モード間の同期が可能となっている。

さらに、我々の音声認識部では、基本周波数の分析による抑揚情報の解析も行なっており、結果が3段階のシンボル系列として得られる。この情報を用いて、例えば肯定の「はい。」と疑問の「はい？」を区別するなど、文字だけでは曖昧な認識結果を正しく解釈することを考えている。(残念ながら、現状では言語解析部がこれに対応できていない。)

3.6 デスクトップTV会議システム

来訪者案内システムのように相手が不特定であるような場合には、かなり状況を限定したとしても、完全な対話システムを作ることは非常に困難である。従ってより実用的な観点からは、システムと人間が柔軟に補完し合えるような仕掛けを組み入れておく必要がある。

このためには対話管理部を洗練するのが本質的な解決方法であろうが、技術的に研究要素が多く残っていることと、それでも完全には解決できないと思われることから、我々は、別途開発中のデスクトップTV会議システム [14] を使って、適宜人間が補助できるような機能を付加した。

具体的には、デスクトップTV会議システムにソケットを使ったプログラムインタフェースを設けることにより、対話管理部との通信を実現している。対話が順調に進んだ場合には、対話管理部から、呼び出して欲しい人の情報が送られ、これを受けたTV会議システムからは、呼び出しの成否と，会議(面会)の終了が通知される。来訪者との対話がうまく進まない場合は[2]、例えば総務課の担当者を呼び出すといったことをすれば良い。

4 今後の課題

我々が重要であると考える課題を以下で述べる。

4.1 モードの効果の把握と機能バランス

先にも述べたように、マルチモーダルインタフェースでは、どのモードにどの様な機能を割り当てるか、あるいは複数のモードの機能バランスをどう取っておくかが重要である [12]。我々は対話管理部に、質問の性質によって入力手段を選択する機能を入れたが、どの様な場合にどのモード(あるいはその複合)を使うべきかという問いに対する明確な知見を得ているわけではない。今後は試作システムの評価を通じ、この点に関する指針を作っていく必要があろう。

4.2 モードの統合追加変更が容易な枠組

例えば音声で「これ」と言いながら画面の一部を指し示すといった機能の実現や、新しいモードの追加等を容易に行なえるような枠組が必要である。一つのアプローチとしては [10] があるが、ジェスチャーによって物を指し示している場合など、そのモードからは無限に続くデータが垂れ流れており、どの時点のデータに意味があるかは他のモードの情報(例えば音声中の指示代名詞の発話時刻)からでなければ特定できないような場合には十分な枠組とは言えない。恐らく、ボトムアップ的な処理とトップダウン的な処理を融合したような枠組が必要である。

4.3 モード統合によるロバストな対話機能

音声で考えると、現状の音声認識ではノイズと発話を区別するために、発話時期を限定したり周囲環境を限定するなどが必要になるが、利用範囲を広げ、より自然な対話を実現するためには、環境に依存しないロ

[2]このような状況をいかに検知するかは今後の課題である。現状では、何度も音声認識に失敗したとか、システム側が同じことを繰り返しているといった程度のことしか分からない。

バストな音声認識技術が必要である。画像認識でも同様のことが言えるが、それぞれの認識技術を向上させるだけでなく、それらを統合することによってロバスト性を上げることも今後重要となろう。例えば画像を使った読唇と音声認識の統合がその一つである。

4.4 ハンドフリーな入力モード

音声をインタフェースに使うメリットは、情報伝達の効率の良さだけでなく、手を自由に使えるという点があげられよう。ポインティングに関しても、マウスやタッチパネルという特定の道具を使うのではなく、離れた場所から指し示したり、大きさをジェスチャーで言えるといったことができれば便利である。我々の試作システムでは、画像認識を頭部位置検出のために用いているが、今後はこのようなジェスチャー認識の機能を取り入れる必要がある。

5 まとめ

本稿では、我々がマルチモーダルインタフェースの研究のために試作している、来訪者案内システムの概要と今後の課題について報告した。試作したシステムでは、コンピュータグラフィックスによる顔アニメーション, 画像認識による頭部検出, 音声認識, およびこれらを統合した対話管理が実現されている。また、案内システムだけでは対処しきれない状況が必ず存在することを考え、デスクトップTV会議システムを組み込むことにより、人間が適宜補助できるような機能を提供している。

マルチモーダルインタフェースの研究には、個々のモードを実現する要素技術をいかにバランス良く利用するかという点が重要である。今後は、試作したシステムに対する評価と新たなモードの追加を繰り返すことによって、バランスの取れた、より実用的なインタフェースの構築を進めて行きたい。

6 謝辞

本試作で用いた音声認識の基本部分は、弊社マルチメディア研究所の音声グループより提供頂いた。ここに感謝する。また、TV会議システムとの結合に協力頂いた関西総合研究所、福永, 藤井両氏に深謝する。

参考文献

[1] R.A.Bolt, "Put-That There: Voice and Gesture at the Graphics Interface", Computer Graphics 14[3], pp.262-270 (1980).

[2] P.Ekman, W.V.Friesen(工藤 訳):"表情分析入門", 誠信書房, 1987.

[3] K.Waters:"A Muscle Model for Animating Three-Dimensional Facial Expression", Computer Graphics, 21(4):17-24, 1987.

[4] 飯田, 有田: "4階層プラン認識モデルを使った対話の理解", 情報処理学会論文誌, Vol.31, No.6, pp.810-821 (1990).

[5] S.Chang, 原島, 武部:"顔の3次元モデルを用いた顔面表情の分析", 信学論(D-II), J74-D-II, 6, pp.766-777 (1991).

[6] 金子, 小池, 羽鳥:"テキスト情報に対応した口形状変化を有する顔画像の合成", 信学論(D-II), J75-D-II, 2, pp.203-215, (1992-2).

[7] 竹内,S.Franks:"表情豊かな顔のアニメーションを目指して", CG シンポジウム'92, pp.169-176 (1992-9).

[8] 塚本, 李, 辻:"合成テンプレートを用いた顔の探索と追跡", 9th Symposium on Human Interface, pp.373-378 (Oct.1993).

[9] 長谷川, 横澤, 石塚:"自然感の高いビジュアルヒューマンインタフェース実現のための人物動画像の実時間並列強調的認識", 信学論(D-II), Vol. J77-D-II, No.1, pp.108-118 (1994).

[10] 島津, 高嶋: "マルチモーダル Definite Clause Grammer (MM-DCG)", 信学論, Vol. J77-D-II, No.8, pp.1438-1446 (1994)

[11] K. Nagao and A. Takeuchi: "Social interaction: Multimodal conversation with social agents", Proc 12th Natl. Conf. on Artificial Intelligence (AAAI-94), pp.22-28 (1994).

[12] 伊藤, 長谷川, 栗田, 速水, 田中, 山本, 大津: "音声・視覚・画像をもつインタラクションシステム", 情処研報, 95-SLP-5, pp.31-38 (1995).

[13] 須崎, 久野, 宮崎, 松山:"顔画像を用いたヒューマンインタフェースの構築", 人工知能学会研究会資料, HICG-9501-5, pp.30-37, (1995-5).

[14] 藤井, 福永, 中井, 宮崎: "インターネット上での多地点間音声通信システムの検討", 信学技報, OFS95-11, pp.17-22 (1995).