音声言語情報処理 22－2
（1998. 7. 24）

視線・音声入力に基づくマルチモーダル・ショートカット機能の提案・試作と評価

○田中 克己
tanaka@krl.toshiba.co.jp
(株) 東芝 関西研究所
〒658-0015 兵庫県神戸市東灘区本山南町 8-6-26

マルチモーダル入力をベースとしたインタフェースとして、マルチモーダル・ショートカットを提案する。これは従来よく行われてきた、自然言語、ジェスチャ入力などの組合わせによる複雑なマルチモーダル入力解析を対象とするものではなく、オブジェクト選択のような単機能をタスクとしている。通常想定される入力モダリティからの入力の不確実性に対して、個々の選択対象をエージェント化し、ユーザとエージェントのリアルタイム・インタラクションにより、ロバストな選択機能を実現するのが特徴である。本報告では視線・音声入力を用いたシステム試作と評価について報告する。

Multi-modal shortcut: a robust selection system using gaze and speech inputs.

KATSUMI TANAKA
TOSHIBA Kansai Research Laboratories.
8-6-26 Motoyama-Minami-Cho Higashinada-Ku
Kobe, 658-0015, JAPAN

This paper presents a real-time object selection system , called multi-modal shortcut, which can deal with gaze and speech inputs with uncertainty. Although there has been many researches which focused on integration of multi-modal information, most of them assumed that each input is perfectly recognized. In addition, real-time interaction with user is an important and desirable feature which few systems have attempted. Unlike those systems, we focused on satisfying these two requirements. In our system, target objects are modeled by agents which react with user's action in real-time. Agent's reaction are based on integration of multi-modal inputs. Our first experiments shows that it is possible to select target object successfully in most cases, even if either of modalities has great uncertainty.

1 はじめに

昨今のヒューマン・インタフェース (HI) 研究の発展にともない、デバイスの開発から具体的アプリケーションに至るまでさまざまな技術的試みがみられる。音声・画像認識技術のような、ユーザの発する自然な入力に基づいたインタフェースは、HI 研究における中心的な課題の一つであるが、現状では以下のような問題がある。

1. 入力の不確実性への対処

   従来のマルチモーダル入力解析方式においては、個々のモダリティがユーザの意図の一部を表現したものであり、複数のモダリティの寄せ集めによりユーザの意図理解という問題解決が行われるという前提がある。しかし、個々の入力がそもそも確実に行えるかという問題点に言及しているものは少ない。

2. リアルタイム・インタラクションの実現

   マルチモーダルインタフェースの典型的な入力である (音声) 自然言語とジェスチャの組み合わせの解析という問題では、自然言語が主モダリティとしての役割を果たし、ジェスチャは参照同定などのための手がかりとして使われるのが一般的である。このようなシステムでは一連のマルチモーダル入力を受付けてから処理することになり、システムからのフィードバックが遅くなる。これはユーザに不快感をもたらす原因となる。

本報告は上記の問題に対する一つのアプローチである。複雑なタスクを取り扱うことを目的としたものではなく、ユーザの注意がどこにあるかを推定し、ユーザによるオブジェクトの選択を支援するというものである。用途としては、GUI 上におけるアイコンやコマンドの選択であるとか、TV におけるチャンネルの選択などが典型的なものである。このような、ユーザによる選択意図を得るための入力情報として、音声認識・画像認識をはじめとした複数のモダリティを用いる。このシステムを、従来スイッチ・キーボードなどにより実現されていたショートカットの発展形として、マルチモーダル・ショートカットと呼ぶ。

上述した通り、マルチモーダル・ショートカットにおいても、また従来のマルチモーダルシステムにおいても最大の問題となっているのは入力の不確実性に対する扱いである。すなわち、音声・画像認識のようなユーザの自然入力を処理対象とするシステムにおいては、完全な認識結果を得ることは不可能である。これは外界の状況 (照明・ノイズ) などによる画像・音声認識の性能変化 (環境による影響) と、また認識結果からだけでは解釈が一意に定まらず、ユーザの利用局面により同一の入力情報であっても解釈結果が異なる (文脈による影響) 点に主に起因している。このような状況においても的確な選択を行うことを目指し、本研究のとったアプローチは、エージェントモデルの導入である。これは個々の選択対象をエージェントとして表現し、エージェントはシステムより得られる結果に基づき、自己の注目度推定・行動を自発的に行うというものである。本報告では、このような目的をもつ選択エージェントの構造についての基本的アイディアをまず説明する。次に、画像処理ベースの視線検出技術・音声認識技術をユーザの入力手段としたマルチモーダル・ショートカット機能を試作したので、その第一次評価を示し、今後の課題について言及する。

2 エージェントモデル

今回とったアプローチを図 1に示す。基本的な流れはユーザ入力の観測、それに基づく意図の推定、行動である。この特徴は、1) 各段階で取り扱う情報を不確実性を持ったまま残し、2) 学習により各情報間の関係を取り扱う点である。図 1で言えば、認識システムより得た認識結果は類似度 (Similarity) という形式で取り扱われ、ユーザの意図推定モジュールに送られる。意図推定モジュールでは、類似度に基づきユーザの意図を推測し、意図それぞれに対する確率に変換して行動決定モジュールに送る。行動決定モジュールでは意図確率から最適の行動を決定し、実行する。各モジュール内での決定は、あらかじめ獲得された学習データに基づいて行われ、学習データは環境の変化に従い適宜更新される。このような方法を取ることにより、不確実で安定性に欠ける入力情報に基づいた場合においても、その場で最適な行動をとるようなインタフェースを構築することが可能になる。

このようなエージェントモデルを用いることの利点は、システムを単純に構成できる点である。一般的にタスクが複雑になるにつれ、認識

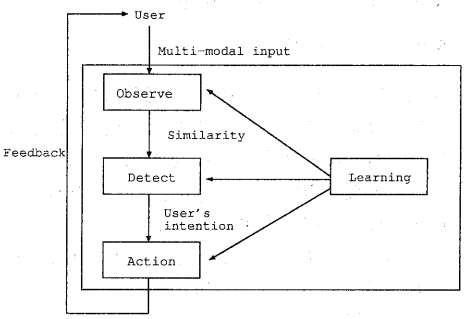

図 1: 基本処理の流れ

システムより与えられる認識結果はモダリティの増加や、モダリティ内の処理の複雑化に伴い増大し、またユーザの意図の種類も複雑化することが予想され、これは特にユーザインタフェースにおいて重要なリアルタイム性を実現する際の大きな障壁となる。このような状況においても、本方式では個々のエージェントがが単純な機能を持ち、単体として動作しているため、基本的にシステムの複雑さはエージェント数(選択対象オブジェクト数)に比例する。

またインタフェースの機能的に見ると、これは音声・視線情報のような自然入力情報とリアルタイム性とを両立させたものであり、従来のような音声自然言語+ジェスチャの一括処理対話とは異なったものである。すなわちタスクは単純であるがキーボード・ショートカットのような素早いインタラクションを実現することが可能になる。このようにマルチモーダル・ショートカットは自然さと素早さを兼ね備えた快適なインタフェースのコンセプトとして位置付けることが可能である。

3 マルチモーダル・ショートカット機能の実現

前章において説明したマルチモーダル・ショートカットの実現を目指した試作として、今回構成したシステムについて説明する。

対象としたタスクは、CRT 画面上でのオブジェクト選択とした。図 2に画面の例を示す。画面上での矩形領域を選択対象であり、選択対象エージェントと呼ぶ。各エージェントの持つ意図は自分が選択されているか否かの 2 種類である。ユーザからのマルチモーダル入力は視線情報・音声情報・マウス・フットスイッチとした。ここでユーザからの生の入力情報はマイクからの音声信号、カメラからの入力画像、マウスからのカーソル位置・ボタンクリック情報、フットスイッチからの ON-OFF 情報となる。

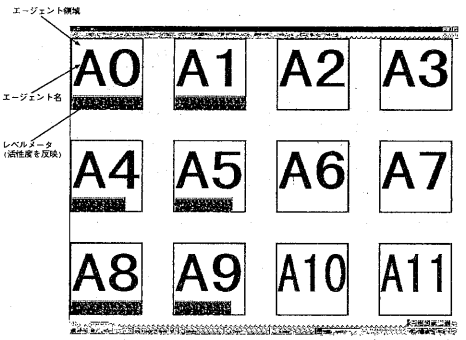

図 2: マルチモーダル・ショートカット システム画面

今回試作したシステムの構成を図 3に示す。ユーザからのマルチモーダル入力はそれぞれの認識エンジンにより認識され、認識結果情報が各エージェントに送られる。各エージェントでは自分の状況(位置・名前など)から各モダリティからの入力を類似度情報に変換する。ここで概念的には図 1に示すように、各エージェントそれぞれが認識機構を持ち、それに応じて処理を行うのがより自然であるが、認識システムをエージェントの数だけ動かすためのコストを考慮すると図 3のように認識システムを共有し、各エージェントでは入力情報の解釈のみを行うのが現実的である。その後、各エージェントでは解釈結果の類似度をもとに意図学習または推定を行う。意図学習モードでは入力類似度情報と意図との因果関係を記録することにより学習を行い、意図推定モードでは記録された因果関係情報と入力類似度情報をもとに意図を推定し、各意図に対する確率を出力する。行動決定部では意図確率をもとにエージェントのとるべき行動を決定し、ユーザにフィードバックを与える。本システムでは図 2中のレベルメータを用いてエージェントが自らの活性度情報をユーザに知らせる。ユーザはそのフィードバックに基づいて次の行動を決定し、それが次のマルチモーダル入力に反映されることになる。上記の一連の処理を定められた時間単位(クロック)に基づき各エー

ジェントで同期して行う。

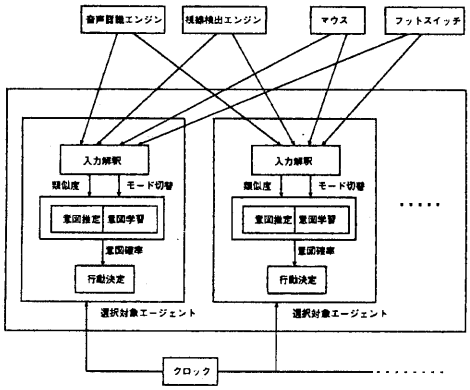

図 3: マルチモーダル・ショートカット システム構成

3.1 モダリティ認識部

音声認識エンジンは、孤立単語認識システムを用いている。これは車載などの耐雑音性能が要求される用途に適した設計がなされており、ノイズが大きい環境でも認識性能が低下しにくいというロバスト性が特徴である。通常音声認識システムにおいては、音声開始区間を高精度で検出するために、発話前にシステムに対しスイッチ等を用いたシグナルを送る (PushToTalk) ことが多いが、今回はインタフェースの自由度を損なうことがないように、そのようにはせず音声信号のみから音声区間を検出することにしている。そのため音声認識システムの一般的な評価手法による測定に比べて認識率の低下が予想される。

視線検出エンジンには開発されたパターン認識に基づく視線検出システムを用いた。これはカメラからの顔入力画像に基づいて瞳の位置などを追跡し、あらかじめ設定された領域に対する類似度を与えることができる。内部の処理はパタン認識の手法に基づいており、設定された領域を見ている場合の顔画像パタン (あらかじめ登録しておく必要がある) と入力顔画像パタンとの照合を行い、両者間の類似度を決定するというものである。

3.2 エージェント内処理

3.2.1 入力解釈

認識エンジンから送られた認識結果の解釈の仕方は単純である。基本的な考え方としては、自分の持つ個別情報 (位置・名前など) と入力情報の類似度を計算すればよい。入力が視線情報の場合には、設定領域に対する類似度から、自分の領域に対する類似度ををを領域間の関係に基づき計算する。また音声認識結果各エージェントの名前に対する認識結果 (スコア) から、自分の名前に対する類似度を (各名前に対するスコアの比として) 求める。

またマウス情報は視線情報と同様に自分の領域に対する類似度を与えるが、これには不確実性は生じないため、各エージェントに対する類似度は 0 または 1 ということになる。この性質を利用して、本システムではマウス情報はエージェントに対する教師信号として、または意図学習・意図推定モードの切り替え (左右のボタンクリックを利用した) として用いる。またユーザによる最終的な選択 (3.2.4に後述) を行うためにフットスイッチからの ON-OFF 情報を用いた。

3.2.2 意図学習・推定手法

今回は意図推定・学習にはベイズ推定の手法を用いた。最も単純な式で表現すれば、各エージェントにとって、入力類似度情報 I が与えられたとき、ユーザが自分を選択しているかどうかという事後確率 P(選択|入力類似度情報 $= I$) は、ベイズの式を用いて

$$P(\text{選択}|\text{入力類似度情報} = I) = \frac{P(\text{選択})P(\text{入力類似度情報} = I|\text{選択})}{P(\text{入力類似度情報} = I)} \quad (1)$$

で与えられる。ここで事前確率 P(選択), 尤度 P(入力類似度情報 $= I$|選択) をあらかじめ与えておくことにより、入力類似度情報 I が観測されたときの事後確率 P(選択|入力類似度情報 $= I$) を計算することができる。ここであらかじめ教師信号が与えられていれば (各エージェントが選択されているかどうかわかっていれば) 学習により事前確率・尤度のデータを得ることができる。

ここでマルチモーダル入力を取り扱う場合は、各入力モダリティ間の因果関係を考慮する必要がある。例えばあるエージェントにとって視線類似度が高いときは一般的に選択意図が高いことが予想されるが、同時に音声認識結果に

基づく類似度が低い場合には、選択意図についての判断は何ともいえなくなってしまう。そこでそのような状況を取り扱えるものとして、本システムではベイジアンネットワーク [1] を採用した。実験で用いたベイジアンネットワークの図的表現を図 4に示す。図中の矢印は因果関係の有無を示すものである。この例でいえば、選択と視線類似度、選択と音声類似度間にはそれぞれ因果関係があるものと考えられる。また視線類似度と音声類似度との間にも因果関係が想定されている (見ているものに対して何か言うという相関が強いと考えられるため)。このような因果関係的知識をあらかじめ表現し、学習・推定により因果関係を考慮した確率的事象が扱えるのがベイジアンネットワークの特徴である。またここでは過去の一致した時間における因果関係を考慮している。これは視線検出結果がほぼリアルタイムで得られるのに対し、音声認識結果が多少のタイムラグを持って得られるという実装時における制約を考慮したものである。

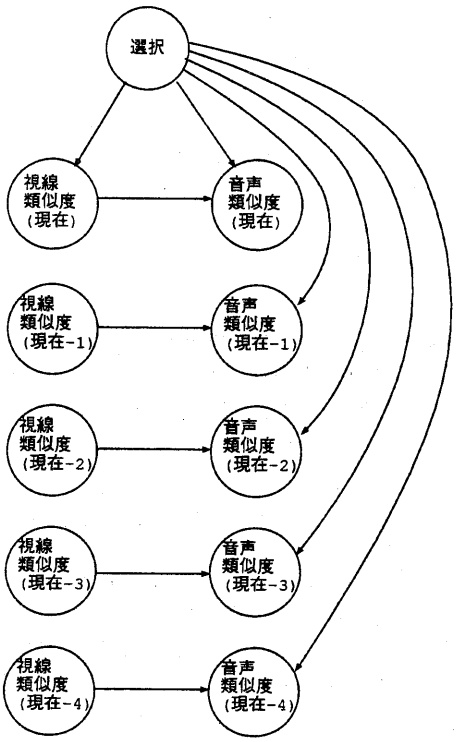

図 4: ベイジアンネットワーク

3.2.3 行動決定手法

前説の手法により得られた選択意図確率に基づいて、各エージェントでは自分が選択されているかどうかを判定し、適切なフィードバックを行う必要がある。ベイズにおける行動決定手法としては、効用理論 (Utility theory) がある [2]。これによると可能な行動 $a_1,\ldots,a_n$ と起こりうる事象 $E_1,\ldots,E_m$ に対してその効用 $u(a_1,E_1),\ldots,u(a_1,E_m),\ldots,u(a_n,E_1),\ldots,u(a_n,E_m)$ が定義される。事象 $E_1,\ldots,E_m$ に対する確率を $P(E_1),\ldots,P(E_m)$ とすると、エージェントの選ぶべき行動は、式

$$E[u(a_i)] = P(E_1)u(a_i,E_1)+\cdots+P(E_m)u(a_n,E_m) \quad (2)$$

を最大にする a_i である (期待効用最大化ルール)。本システムで取り扱われるエージェントにしてはとるべきフィードバック行動は基本的に (活性化・沈静化) の 2 種類であり、事象は (選択, 非選択) の 2 種類である。ここで効用テーブルを図 5のように構成し、意図推定部より与えられる選択確率 P(選択), P(非選択) と式 (2) を用いることにより、各エージェントにおける最適行動を決定することができる。本システムでは実際に図 5のように効用テーブルに具体的な値を与えている。これはエージェントがユーザの意図を正しく反映した行動をした場合には 1, しなかった場合には -1 を与えるという最も単純な設定である。

フィードバック \ ユーザの意図	選択意図あり	選択意図なし
活性化	1	-1
沈静化	-1	1

図 5: 効用テーブル

この手法を用いることによりエージェントの行動は決定できることになるが、実際にはマルチモーダル入力は不確実情報でありそれに基づいた意図推定を行っている以上、推定も不確実にならざるを得ない。それはフィードバックの誤り (選択意思がないのに活性化する、またはその逆) となる。これを防ぐための手法として、今回は各エージェントにおける効用を時間により累積して、フィーバックはその累積値により決定することによりユーザインタフェースとし

てのロバスト化を図っている。

3.2.4 ユーザによる最終選択行動

最終的にエージェントを選択するのはユーザである。エージェントは活性度によるフィードバックを与えるだけで、自分が選択されているかどうかを間違いなく決定することは不可能である。これは不確実な情報を入力として用いる際には必ず生じる問題である。今回はその取り扱いを、エージェントのフィードバック状態と簡単な確実情報入力モダリティ (ON-OFF スイッチ) により行う。選択行動のための手続きは非常に単純であり、各エージェント内において以下のように行われる。

1. スイッチが押された時刻に、累積効用が上限値に達していない (死んだ) 場合は選択意図なしと判断し、以後の意図推定・フィードバックは行わない。
2. 累積効用が上限値に達している (生き残った) 場合は以後も同様の意図推定・フィードバックを続ける。
3. このステップを生き残りのエージェントが一つになるまで続ける。

これにより候補の絞り込みが順次行われ、システムとして基準時刻ごとに選択行動を繰り返す。ユーザはエージェントの累積効用 (直感的に活性度を意味する) をフィードバックにより簡単に確認できるため、スイッチにより選択を決定する行動も負担なく行うことができる。

今回は選択をスムーズに進めるための工夫として、エージェントの候補の絞り込みが行われた時にエージェント間の関係を、よりユーザが選択しやすいように変化させることにした。具体的には、視線での選択が容易になるように、互いの位置関係を、エージェント間の距離をなるべく離すように調整した。

4 実験と評価

今回作成したシステムの基本性能を確認するための実験を行った。図2のように規則的に配置されたエージェントが提示され、ユーザがマルチモーダル入力を用いて意図したエージェントを選択することをタスクとする。

具体的な実験の手順を以下に示す。

1. ユーザはマルチモーダル入力を用いた意図学習を行う。これはマウスを用いて選択意図対象のエージェントを指定し、視線・音声入力情報をシステムにあらかじめ与えておく教師付き学習により行う。
2. すべれのエージェントに対して学習を行なった後、ユーザは学習を行ったエージェントの選択を、マルチモーダル入力により試みる。具体的には、学習時の行動 (視線による注視または名前の発声) をそのまま再現するように試みた。
3. システム側では、上記の間のユーザのマルチモーダル行動、選択行動が達成されたかどうかといった情報を記録し、システムの性能評価のためのデータとする。

ここで基本的な評価基準として、ユーザが意図したエージェントを正しく選択できたかどうか (選択成功率) が重要である。次に選択に成功したとして、それまでに要した時間 (選択時間) が評価の対象となる。

考慮すべきパラメタはタスクの困難度を反映していなければならない。今回のような選択タスクでは、選択対象のエージェント数がそれにあたる。エージェント数が増加すると、視線検出・音声認識において判別候補が増加するため、入力情報の不確実度が増大する。今回は図2のように、4x3 に配置した 12 エージェントと、同様の配置で 8x6 に配置した 48 エージェントの選択という 2 種類のタスクを準備した。

4.1 実験結果

エージェントの配置をパラメタとした結果を実際の視線検出・音声認識性能とともに表1-2に示す。比較のため同一システムで視線検出のみ、音声認識のみを用いた場合の結果を付記する。

これより、以下の考察を導くことができる。

1. 音声のみ・視線のみの場合よりも音声、視線を両方使用したほうが選択成功率が高く、また一例を除いては選択時間も短い。これはマルチモーダル情報を使用した効果を示すものである。この原因としては、視線検出システムによりリアルタイムに類似度情報を得ることが可能なため、それがたとえ精度が低いものであってもデータ量の豊富さにより学習が容易であり、統計的には有意に精度の高い予測を行うことができているためと考えられる。これはエージェントの絞り込みにとっ

表 1: 選択性能 (48 agents(8x6 に配置))

項目 \ モダリティ	選択成功率	選択時間 (平均)	音声認識結果 (再現率 / 正解率)	視線検出結果 (再現率 / 正解率)
音声のみ	77.6 %	9.9 秒	77.4% / 77.4 %	-/ -
視線のみ	63.3 %	22.0 秒	-/ -	59.0% / 9.7 %
音声＋視線	95.8 %	9.0 秒	84.6 % / 84.6%	62.1% / 6.9 %

表 2: 選択性能 (12 agents(4x3 に配置))

項目 \ モダリティ	選択成功率	選択時間 (平均)	音声認識結果 (再現率 / 正解率)	視線検出結果 (再現率 / 正解率)
音声のみ	83.3 %	8.1 秒	84.2% / 84.2 %	-/ -
視線のみ	83.3 %	12.0 秒	-/ -	49.4% / 31.8 %
音声＋視線	100 %	6.4 秒	83.3 % / 83.3%	69.1% / 38.4 %

てたいへん有益な情報であり、音声認識システムのように断続的ではあるが比較的精度の高いシステムとの組合わせによる相乗効果が高いものと思われる。

2. 音声認識・視線検出の精度と比較して、本システムにおいての選択成功率は一例を除き上回っている。これは認識性能が低いモダリティの寄せ集めでも高い精度を持つ認識インタフェースの実現が可能であることを示唆している。これは視線検出のみを用いた場合には精度が悪くなるが、今回用いたエージェントの段階的な絞り込みと位置調整によりある程度の選択が可能であることからも裏付けられる。

上記の考察より、基本的には視線・音声を両方使用したマルチモーダルシステムの優位性をうかがうことができる。

4.2 失敗原因の分析

次に失敗原因について分析する。本システムを用いて選択に至らなかった場合は、以下の2例に分類される。

1. 予測の失敗
   意図学習・推定段階で、正解エージェントの選択確率が低く押さえられる場合に相当する。これは主に学習データ不足による sparse data problem によるものであり、特に音声認識時に多く発生する。
2. ユーザ選択の失敗
   エージェントの効用を計算し、正解エージェントにとってそれが上限値に達するところまではうまく動いているが、そこでユーザが確認のためのスイッチ (フットスイッチを用いている) を押し、システムが検出するまでの間に効用値が上限を下回ってしまい、意図した選択が行われない場合である。これは人間・システム間ののインタラクション時のタイムラグに起因するものと考えられる。今後の改良において、システムのリアルタイム性とロバスト性を両立させる手法には工夫が必要である。

5 関連研究

従来のマルチモーダル・インタフェースとしては、ユーザによる音声・ジェスチャなどの入

力を、たいていは一発話文単位で一括して処理し、一括して応答を返すシステムが主流であった[3, 4, 5]。本システムは、これは視線という、連続的な入力情報が得られるモダリティを用いたことにより、ほぼリアルタイムの入力に対してリアルタイムに応答を返し続けるという点でマルチモーダル・インタフェースの新たな形態を示している。すなわちユーザによる意識的な入力(言語・ジェスチャなどの)に加え、視線のような無意識的入力情報を受け付けることが可能になるからである。今回の実験ではそのような連続的・無意識的入力が、たとえ精度の低いものであっても選択というタスクにとって有効に働くことを示している。このようなリアクティブで自由度の高いインタラクションを実現するために、マルチエージェント的なシステム構成を採用しており、それは環境が変化することが容易に予想される今後の実使用においてさらに有効となることが期待される。

視線ベースのインタフェースとしては、主に障害者の入力補助を目的とした研究が行われている。このような研究では視線検出システムに高い精度のものを用い、視線入力キーボード、視線マウスなどの高度なタスクを実現しているものが存在する[6]。ただこのような視線検出システムは頭部の固定、特殊眼鏡装着などの制約があるためユーザへの相当な負担となり、また実現コストが飛躍的に高いため、一般的なインタフェースとしては実用的とは言えない。低精度でタスクを実現する今回のアプローチのほうに一般性が存在する。

6 おわりに

本報告では、マルチモーダル・ショートカット機能のコンセプトを提示し、視線検出・音声認識システムを使用した実現と基本性能の評価を行った。結果は12個のエージェントのの選択タスクにおいて100%,48個においては95.8%を、比較的精度の低い認識システムを用いた場合に達成できた。これは将来の実用化に向けて有望な結果が得られたと言える。特に視線検出は無意識的情報がリアルタイムで得られるという新規な特徴をもち、音声認識のような従来のマルチモーダルシステムで用いられている認識システムとの組合わせが非常に有効であることが確認できた。

今回の基本試作・評価をふまえ、PC用インタフェースを当面のターゲットとして、実アプリケーションに即したマルチモーダル・ショートカット・システムの実現を進めていく。またこのような視線検出・音声認識に基づくインタフェースの非接触性、快適性を活用できるアプリケーションの発掘を行っていく。

参考文献

[1] David Heckerman. A tutorial on learning with bayesian networks. MSR Technical Report MSR-TR-95-06, 1995.

[2] 繁桝算男. ベイズ統計入門. 東京大学出版会, 1985.

[3] R.A. Bolt. Put-that-there: Voice and gesture at the graphics interface. *Computer Graphics*, Vol. 14, No. 3, pp. 262–270, 1980.

[4] Philip R. Cohen. Natural language techniques for multimodal interaction. 電子情報通信学会 論文誌, Vol. J77-D-II, No. 8, pp. 1403–1416, 1994.

[5] Philip R. Cohen et al. Quickset: Multimodal interaction for distributed applications. In *Proceedings of the Fifth Annual International Multimodal Conference. Association for Computing Machinery*, 1997.

[6] Gregory Edwards. A tool for creating eye-aware applications that adapt to changes in user behavior. In *ASSETS 98*, p. (to appear), 1998.