ヒューマンインタフェース 65－4
音声言語情報処理 10－4
（1996．2．29）

クリックを使わないマウスの動きと音声を入力とするインタフェース

中嶋秀治　加藤恒昭
{nakajima，kato} @nttnly.isl.ntt.jp
ＮＴＴ情報通信研究所
〒238-08　神奈川県横須賀市武1-2356

「これをここに置いて」のような指示語等を含む音声とマウスカーソルの動きからなる入力によって，クリックを用いずに指示語に対応する計算機ディスプレイ上の対象を指示できるマルチモーダル・ユーザ・インタフェースを提案する．本インタフェースは，マウスのカーソルの移動速度の変化と指示対象領域への入／出から指示対象候補でのカーソルの滞在時間を抽出し，音声から指示語の発声された区間を抽出し，これらの重なりが最大になるように対応付ける．評価実験の結果，指示対象と指示語との対応付けの正解率は９３．９％であり，本インタフェースで用いた，指示対象と指示語との対応付け手法の有効性が確認された．更にMultimodal WWW Browserを例として本インタフェースの応用可能性の考察を行なう．

An Interface using Mouse Movement and Voiced Command

Hideharu NAKAJIMA　and　Tsuneaki KATO
{nakajima，kato} @nttnly.isl.ntt.jp
NTT Information and Communication Systems Laboratories
1-2356 Take，Yokosuka-shi，Kanagawa，238-08 JAPAN

This paper proposes a new multimodal user interface employing deictic words and mouse movement. After the user points to objects using the mouse cursor without clicking and gives a verbal command such as "Put this here," this interface associates the objects with the deictic words one by one. This proposed interface detects the desired candidates based on the slowing of the cursor movement while the cursor is on each object. Next, the interface identifies deictic words in the speech sound wave taken from the verbal command. Finally, it associates the objects to the deictic words. The association procedures for this interface are proven through experimentation to successfully output more than 90% correct object-deictic word combinations. Moreover, a Multimodal WWW Browser is discussed as one application of this interface.

1．はじめに

音声や動作等の複数のモダリティの相補によって，単一のモダリティでは煩雑となる情報入出力を効率的・効果的に行なうというマルチモーダル・ユーザ・インタフェース（MUI）の重要性が指摘されている[1]．本稿では，ユーザが対座している計算機ディスプレイ上のグラフィカル・ユーザ・インタフェース（GUI）を発展させたMUIの１つとして，「これをここに置いて」等の音声と，クリックを使わない指示動作からなる人間のマルチモーダル発話を理解する（指示語と対象とを対応付ける）MUIを提案し，その得失と応用可能性について考察する．

MUIの一般的な利点は，文献[1]で指摘されているように，GUIへの直接操作（Direct Manipulation ）と自然言語入力とを融合しGUIをMUIに発展させることによって，それぞれの欠点を補い，効率的かつ効果的なインタフェースを実現できることである．例えば作図の場面では，複数の線の色を１回の命令で変更することができるなどの効率のよい命令入力をMUIによって実現できることが指摘されている[2]．このような意味で，MUIには次のような入力機能の実現が期待される．

・１回の命令で１つの処理内容を複数の対象に対して実施できる機能（これとこれを捨てる），
・１回の命令で複数の処理内容を１つの対象に対して実施できる機能（これを拡大して黄色にする），
・１回の命令で複数の処理内容を複数の対象に対してそれぞれ実施できる機能（これを捨ててこれを拡大する）

一方，マルチモーダル対話では電話対話に比べて「これ」「ここ」等の近称の指示語が多く利用されるという分析結果が報告されている[8][9]．またユーザは近称の指示語と指示動作を使うことによって，「タイトルバーの左端の」のような正確な位置を示す修飾語句を使わずに簡潔に発話できると考えられる．従ってMUIは「これ」や「ここ」に対応する対象や位置を指示動作から明確にすることが必要となる．

従来のMUIの研究は，ユーザが指示している位置をユーザの指示動作だけで明示できる研究と，指示動作だけでは明示できない研究とに分けることができる．前者のMUIの例として，位置をマウス・クリックの位置で明示するもの[6]と，タッチパネルへの接触した位置で明示するもの[3][7]がある．これらのMUIでは，ユーザが対座している計算機のディスプレイ上に指示対象が置かれている．一方，後者のMUIの例として，Bolt[4]や福本ら[5]の研究がある．この２つの研究では，指示対象がユーザから離れたスクリーン上に置かれており，タッチを指示に使えない．そこで，指示動作を含む腕や指先の連続した動作と，指示語の発声との時間的相関関係を利用して指示対象を判定している．このような方法を，ユーザが対座している計算機ディスプレイ上のGUIまたはMUIに適用した例はない．そこで，この方法を本MUIに適用する．

以下ではクリックを使わない指示動作と音声とを入力とするMUIにおいて指示語と指示対象とを対応付けるアルゴリズムを決めるために行なった実験を２節で述べ，３節で本MUIの構成を述べる．そして本MUIを使って，数名の被験者が行なった発話の中での指示語と対象物とを対応付ける実験について４節で述べ，５節では本MUIの処理過程の考察と応用可能性を述べる．

2．予備実験と設計方針の決定

2．1　実験

本MUIの設計のために次の予備実験を行ない，クリックを使わないマウス・カーソルの動きによる指示動作と音声からなる発話の観察を行なった．

この予備実験で被験者に提示した画面の例を図１に示す．図１のように，画面に１辺が1.5cmの正方形を81個（縦9列横9列），相互の間隔を1cmとして配置した．その中から乱数により3つの正方形を選び，被験者に指示させた．被験者に指示させる正方形にはそれぞれピンク，青，緑の色を付け，指示

させない正方形は褐色にし，更に図１のように指示語と指示対象の色とを対応付けて表示した．また，Work Station （以後WS）の光学式マウスのカーソルを丸型にし，マウスのPointer Control のaccelerationとthresholdは被験者によらず一定値とした．

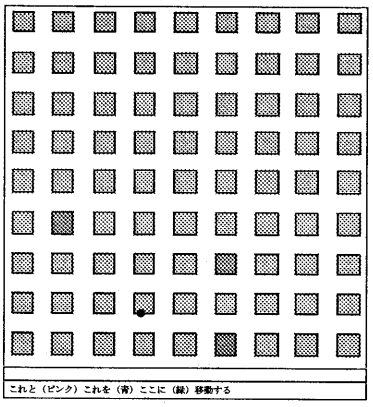

図１　被験者に提示した画面の例

被験者には「これとこれをここに移動する」と言いながら画面上のそれぞれの正方形を指示するよう教示した．これを１回の発話と数える．指示する正方形の組み合わせは１人当り10通りで，すなわち１人当り10回発話した．特に教示を与えず，6名の被験者に自由に発話を行なわせた．

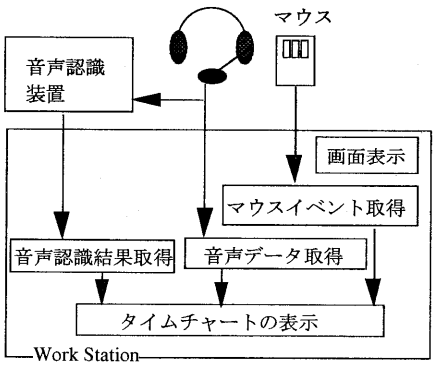

図２　データ収集系

データ収集系を図２に示す．実験では，まず図１の画面をディスプレイに出力する．その後，ユーザの発話毎にマウス・イベント（位置座標，正方形へのEnterとLeave），そのイベントの発生時刻，及び音声データをWSに記録した．また，図２のようにマイクからの出力を単語音声認識装置に通し，認識結果をRS-232Cを通してWSに記録した．実験の間，画面上でのカーソルの移動の様子を8ミリビデオテープに記録した．

WSに格納した各データの時間情報を用いて，チャートを作成した．その１例を図３に示す．

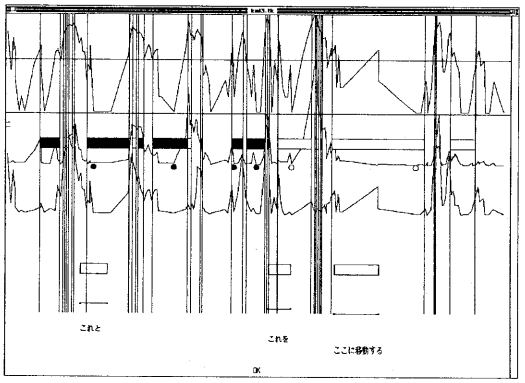

図３　音声とマウス動作の時間関係の１例

図３の横軸は時間である．3本の折れ線グラフはカーソルの速度変化（下2つはx成分とy成分の速度変化）を示す折れ線グラフである．上側の横1列に並んでいる各長方形はカーソルが図１の各正方形に滞在していたそれぞれの時間帯である．以後，滞在時間帯と呼ぶ．長方形の左端の時刻に図１の正方形のどれかにEnterして，長方形の右端の時刻にLeaveしたことを示している．下側の3つの長方形は音声データの短時間平均エネルギーが一定値以上かつ有声化確率が一定値以上となった時間帯である．更に，最下行の文字列は音声認識結果の文字列である．

２．２　結果

全被験者の全データを図３のように表示し分析した結果，以下が観察された．指示のときに6名全員（60発話，180回の指示）が正方形の領域に入って指示を行なった．また，速度が一定値以下になった場合を「停止」とすると，6名全員が指示対象の正

方形の中でカーソルを停止した．その他，動作の様子には，指示対象上でクリックする動作，指示対象を囲むように動いた後に正方形の中で停止する動作，移動してきて指示対象の中で停止する動作が見られた．

一方，WSに格納された音声の短時間平均エネルギーから，音声は「これと」「これを」「ここに移動する」のように文節に近い単位（以後，文節と呼ぶ）で発声される場合が51発話（85%）存在した．このような単位に分かれる理由としてマウスの操作性や指示される対象の画面上での位置による影響が推察される．更に文節の発声時間帯と指示された対象内での滞在時間帯はほぼ同期していた．そして，発声時間帯と指示された対象内での滞在時間帯との重なりは，発声時間帯と指示されていない対象内での滞在時間帯との重なりよりも長かった．

以上の結果から設計方針を以下の通りに決めた．すなわち，指示された対象の候補の抽出では，マウスのカーソルの移動速度が一定値以下になった時に滞在していた領域に対応する対象を指示された対象の候補として抽出する．また，指示語と指示された対象の候補との対応付けには，発声時間帯との間で滞在時間帯の重なりが最も大きくなる対象を対応付ける．

3．本MUIの構成と処理内容

全体の構成を図4に示す．本MUIは図2と同様の単語音声認識装置とWSで構成されている．MUIは，図2と同様に，画面上に指示対象の描画を行い，それらに対するユーザのマウス操作で発生したマウス・イベントと時刻の取得を行い，指示対象候補を抽出する．音声は音声認識装置に入力され，認識結果がRS232Cを通してWSに転送される．この認識装置には2節の文節に相当する語を登録した．また，WSのオーディオ・ポートから取得した音声から，認識装置で認識された語の発声時間帯（指示語の位置）が抽出される．その後，マウスの動きから抽出された指示対象候補と音声中の指示語との対応付けが行なわれる．

次に，各部の処理を述べる．WSで録音された音声から短時間平均エネルギーと有声化確率がともに一定値以上になる時間帯を計算し，その時間帯を「これと」のような文節が発声された位置として抽出する．以後，この区間を文節時間帯と呼ぶ．文節時間帯は音声認識結果と順に対応付けられる．

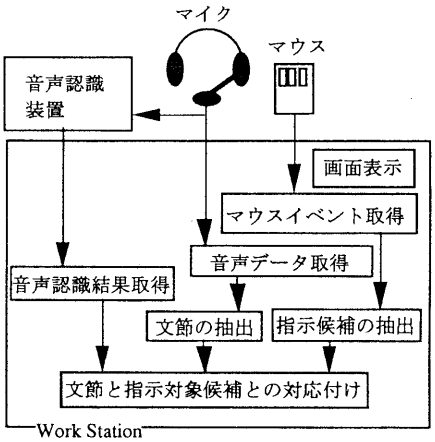

図4　インタフェース全体の構成

取得されたマウス・イベントとその時刻から，マウスの移動速度が一定値以下である時に滞在していた領域を，指示された候補と判定する．以後，指示された候補へのEnterからLeaveまでの時間帯を指示対象候補滞在時間帯と呼ぶ．

文節時間帯と時間的に重なりがある指示対象候補滞在時間帯のうち，その重なりが最も大きい指示対象候補滞在時間帯を持つ指示対象候補を指示対象と判定する．

4．評価実験

2節の予備実験と同じ要領で評価用データの収集を行なった．ただし，

R1：指示の際に指示対象の領域へマウス・カーソルを入れること，

R2：指示動作と移動の間で速度に違いをつけること，

R3：発声と指示とのタイミングを合わせること

を教示した．被験者は予備実験に参加した被験者の中の5人とした．全部で50発話，150指示が行なわれた．

評価のために表１の4通りについて本MUIの正解率の平均値を計算した．ただし予備実験データは評価実験に参加した被験者5人のデータを用いた．表１の全データの正解率の平均値は，全データ50発話（150指示）を対象にして計算された値である．文節発声のデータの正解率の平均値は，全データのうち，文節に１つずつ指示語が含まれるような発声が行なわれた発話（予備実験データでは42発話（126指示），評価実験データでは44発話（132指示））だけを対象にして計算された値である．

表１　正解率の平均値

	予備実験	評価実験
全データ（%）	(1)　79.3	(3)　86.7
文節発声のデータ（%）	(2)　86.4	(4)　93.9

５．考察と応用可能性

表１のように，ユーザが入力の際にR1からR3のような教示を意識することなく入力した場合の正解率は79.3%であった．今回は明示的に教示しなかったが，文節発声を教示した場合には，(1)から(2)や，(3)から(4)のように，正解率の向上を期待できる．(1)から(3)や，(2)から(4)では，指示動作だけから出力される指示対象候補の数の減少が見られた．従ってR1からR3に加えて文節発声を教示すれば，(4)の正解率が期待できる．従って，本MUIでの指示対象抽出と，指示語と指示対象との対応付け手法は有効である．このように本MUIでは，マウス・クリックを使わずにマウス・カーソルの動きだけによる指示動作が可能となった．

また本MUIは，動作データを取得し認識する特殊な装置を必要としない．

次に，応用可能性とクリックを使わずに指示できる利点を考える．

GUIの画面には，クリックしたりタッチしたりすると何等かの処理が開始されるボタンのような物体がしばしば表示される．それらの中には，アフォーダンス[10]への配慮から「押す（クリックする，タッチする）となになにが始まってしまう」ということを一層ユーザに感じさせる物体もある．このような物体を含むGUIで，指示を行なう場合にクリックやタッチを使うと，指示動作としての意味とGUIへの直接操作である「押す」の意味との間で曖昧性が生じる．前者の指示動作として，モディファイ・キィーとクリックとの併用などの回避策が考えられるが，煩雑となる．従って，どの物体が指示されているのかを，クリックを使わずに行なわれた指示動作と音声を元に認識する機能は必要である．従来のMUI研究では，指示される対象がボタンのような物体ではないので問題にはされなかった．

以上のように利点は，モード切り替えやモディファイ・キィーを併用することなく，ボタン等を指示する場合に利用できることにある．

１つの応用例として，本MUIの入力技術の，Internet上のWWWの文書を閲覧するためのMosaic等のブラウザへ適用がある．ブラウザの画面に表示されるアンカーと呼ばれる文字列は，ボタンと同様に機能する．つまり，それをクリックすると別の文書が表示されたり，文書の取得（ftp）が開始されたりする．例えば，会議情報のページには会議場付近の地図や会場の地図のような，研究者のホームページにはこれまでに書かれた論文やその抄録のような，ftp可能な文書が複数個置かれている場合がある．

それらの内の２つの文書をftpして取る場合を考える．従来のブラウザでは，ユーザが１つ目のアンカーをクリックして，取得が完了するまで待ち，完了したらユーザが２つ目のアンカーをクリックして，取得が完了するまで待たなければならない．つまり待ち時間が分散して発生し，ユーザは２つ目のアンカーをクリックする為に，１つ目の処理の終了を検

知せねばならないので不便である．しかし，１回の命令で「これとこれを取ってくる」という音声とマウスカーソルの動きを使って対象を指定し，ftpできれば，分散して発生していた待ち時間を１つにまとめることができ，利便性が向上すると考えられる．クリックを従来と同じ意味で使えるので，ブラウザをMultimodal WWW Browserに発展させても，従来のブラウザの利用感覚が損なわれることはない．

既に，ブラウザのマルチチャネル化は試みられている[11]．これはアンカーをクリックする代わりに，アンカーの言葉を発声することによってページの切り替え等を行なうものである．しかし，１節で述べたMUIの利点は実現されていない．

また，「これ」「ここ」に相当する対象への指示動作は，ボタンを「押す」とかスライドバーを「摘んで動かす」ような直接操作と比べると，別の分類に属する操作であると考えられる．このような操作は，クリックやタッチを使わずに実現できる方が望ましいのではないだろうか．実際，Apple社のMacintoshやMicrosoft社のWindowsでそのような区別が見られる．これらのアプリケーションでは，それを利用する時に用いるマウスのドラッグやクリックのような本来必須の動作と，Balloon Helpを出すためにカーソルをそれぞれの領域に入れるような動作とが区別されている．

このようにクリックを用いない指示動作と音声とを使った入力技術は，Help機能を充実する，Help機能を対話化するという意味でも存在意義がある．

６．おわりに

本稿では，GUIのMUI化の観点から，マウス・クリックを使わない指示動作と音声入力中の指示語との対応付けを行うMUIを提案した．そして，その利点として，現在のGUIの直接操作感を損なうことなく，指示としての入力が可能となることを挙げた．本MUIの実装では，音声認識装置として日本電気製の大語彙音声入力装置DS-1000を，音声処理にはEntropic社のESPSを利用した．本MUIはSparcStationIPX上で動作している．

［参考文献］

[1]Cohen, P.R.:"Natural Language Techniques for Multimodal Interaction",信学論 vol.J77-D2 No.8, P.1403-1416, 1994.

[2]Hiyoshi M. ,Shimazu H.:"Drawing Pictures with Natural Language and Direct Manipulation",Proc of the Coling'94, P.722-726, 1994.

[3]安藤,北原,畑岡: "インテリアデザイン支援システムを対象としたマルチモーダルインタフェースの評価",信学論 vol.J77-D2 No.8,P.1465-1474, 1994.

[4]Bolt R.A.:"Put-that-there:Voice and gesture at the graphics interface",ACM Computer Graphics,14, 3, P.262-270,1980.

[5]福本,間瀬,末永:"動画像処理による非接触ハンドリーダ",第7回ヒューマンインタフェースシンポジウム,P.427-432, 1991.

[6]井本,山田,嵯峨山:"HMM-LR方式音声認識サーバを用いたマルチモーダル入力",SD-9-6,信学全大,1995.

[7]Loken-Kim, K.H.:"翻訳通信環境におけるマルチモーダル入力の分析と統合",情処研報SLP7-12,1995.

[8]加藤,中野:"マルチモーダル対話における対象物の同定について",第50回情処全大4R-5,1995.

[9]谷戸,キュンホ,ファイス,森元:"道案内タスクにおけるマルチモーダル対話の会話文の特徴分析",信学論vol. J77-D2 No.8,P.1475-1483,1994.

[10]Norman D.A.:"テクノロジー・ウォッチング",P.28,新曜社,1995.

[11]土肥,石塚:"自然感の高い擬人化エージェントとWWW/Mosaicの結合",第51回情処全大6U-4,1995.