これをここに置いて：バイモーダルインタフェースの提案

6U－2

中嶋秀治　加藤恒昭
ＮＴＴ情報通信研究所

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インタフェース（MMI）の重要性が指摘されている[1]．本稿では，このようなMMI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MMIを提案し，その得失について考察する．

２．関連研究と本システムの位置付け

指示動作を伴った「これをここに置いて」のような言葉を入力として「これ」や「ここ」に対応する対象や位置を明確にするMMIでは，指示動作からそれらを明確にすることが必要となる．従来の研究は，このような指示位置をユーザーが明示するMMIと明示しないMMIとに分けることができる．

まず，前者のMMIには，例えばHiyoshiら[2]，安藤ら[3]があるが，それぞれ，マウス・クリックされた位置，パネルへ接触した位置をもとに，選択された指示対象を判定している．これらのMMIでは，ユーザー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，後者の指示位置を明示しないMMI，例えばBolt[4]や福本ら[5]，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，マウス・クリックやタッチを指示に使っていない．その代わりに，腕や指先の動きによる連続した指示動作から，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．

本稿で提案するMMIでは，前者のMMIと同じように，ユーザが対座している計算機ディスプレイ上に置かれた対象を指示する状況を扱う．しかし，指示対象の判定と指示語との対応付けには，後者のMMIの考え方を用いる．つまり，クリックやパネルへのタッチを使わずマウスカーソルの動きから指示対象候補を抽出し，指示語との対応付けを行なう．

３．インタフェースの構成

３．１　予備実験と設計の方針の決定

本MMIの設計のために次の予備実験を行ない，発話を観察した．実験では，画面に８１個の正方形を縦９列横９列に配置し，そのうちランダムに３つの正方形を１０通り選び指示させた．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３つの正方形を順に指示させた．光学式マウスのカーソルを丸型にし，指示の仕方については特に教示を与えず，６名の被験者に自由に発話を行なわせた．実験の間，マウス・イベント（位置，領域へのEnterとLeave）と時刻と音声をWork Station（WS）に，画面上でのカーソルの移動の様子を８ミリビデオに，それぞれ記録した．

実験結果から以下が観察された．指示のときに６名全員が正方形の領域に入って指示を行なった．また，平均速度がある値k以下になったときを「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．また，動作として，指示対象をクリックする動作，指示対象を囲むように動いてそ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，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多かった．これにはマウスの操作性や対象の画面上での位置が関係すると推察される．更に文節の発声時間帯と指示対象内でのカーソルの滞在時間はほぼ１対１に対応していた．

以上の結果から以下の設計方針を決めた．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k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．指示語と指示された対象の候補との対応付けには，文節の発声と指示動作との１対１対応を仮定し，時間的相関関係を利用する．

３．２　本MMIの構成と処理

全体の構成を図１に示す．本MMIは単語音声認識装置とWSで構成されている．

MMIは，画面上の指示対象の描画を行い，それらに対するユーザー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．この認識装置には文節に近い大きさの語を登録した．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区間（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．

各部の処理の詳細を述べる．

WSで録音された音声から有声化確率と短時間平

Put this here：A bi-modal interface.
Hideharu NAKAJIMA and Tsuneaki KATO,
NTT Information and Communication Systems Labs.
1-2356 Take, Yokosuka-shi, Kanagawa 238-03, JAPAN
e-mail : {nakajima, kato}@nttnly.isl.ntt.jp

均エネルギーを計算し，両者の重なりが有る区間を「これと」のような音声中の文節位置として抽出する．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．

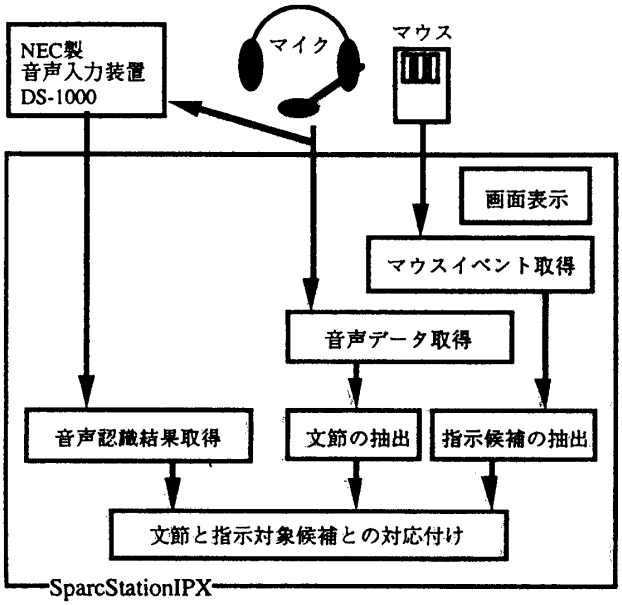

図1　提案するインタフェース全体の構成

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k以下になった時に滞在していた領域を，指示された候補と判定する．ここで，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

3．1の実験と同じ要領で評価実験を行なった．ただし，指示の際に指示対象の領域へマウスのカーソルを入れること，指示動作と移動との間で速度に違いをつけること，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5人で，全部で50発話，150指示が行なわれた．

3．1で述べたように，設計方針として文節の発声を仮定したが，本評価実験で収集した50発話の内，それを満たす発話は44発話であった．この44発話の132指示についての，本MMIが出力した指示語と指示対象の組の正解率は，93.2%であった（表1）．

表1　正解率

被験者	A	B	C	D	E	total
正解数	26	30	15	24	28	123
全出力数	27	30	15	30	30	132
正解率（%）	96.3	100	100	80.0	93.3	93.2

5．考察

小規模な評価実験ではあるが，本MMI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が有効であることが確認された．本MMIでは，マウスのクリックを使わずにマウスの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MMIには動作を認識する特殊な装置を必要としない．

次にクリックを使わずに指示できる利点を考える．

例えば，オーディオパネルのようなGUIに，そのGUIのボタン等の機能に関しての，指示動作と音声による，質問を受け付けるヘルプシステムを付加する場合を想定する．本来，そのGUIは，ボタンを押せばテープの再生・巻き戻しが始まるといった操作の直接感をユーザに与える．この場合クリックにはpushの意味がある．しかし，付加したヘルプシステムで，例えば「これを押すとどうなるの？」という質問に伴う指示動作にクリックを採用した場合，この指示のクリックと，本来の操作のためのクリック（pushの意味）との区別が付かず曖昧性が生じる．後者の指示動作として，shiftボタンを押しながらのクリックを使うなどの曖昧性の回避策も考えられるが，煩雑となる．それに比べて，後者の指示動作として，クリックを使わずに指示できる本MMIの手法を用いる場合には，煩雑にならない．

一方，課題として，音声が文節に近い単位で発声されるという仮定を除いた実装が挙げられる．指示対象間の距離が近いことによって仮定が崩れた場合には正解率が下がるからである．

6．おわりに

本稿では，マウスの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MMIを提案した．音声認識装置には日本電気製の大語彙音声入力装置DS-1000，音声処理にはEntropic社のESPSを利用した．本MMIはSparcStation上で動作している．今後は，指示語と指示対象候補との対応付けの高度化，様々な指示動作への対処，MM対話システムへの適用を行なう予定である．

[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. et al.:"Drawing Pictures with Natural Language and Direct Manipulation", Proc of the Coling'94, P.722-726, 1994.

[3]安藤: "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 他: "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.