ユーザの注目情報を利用したマルチモーダル・インタフェース

茶園 篤*，長田 誠也**，國枝 和雄*
*NEC ヒューマンメディア研究所
〒630-0101　奈良県生駒市高山町 8916-47
Tel. 0743-72-3641　FAX. 0743-72-3599
E-mail：chaen@hml.cl.nec.co.jp
** NEC C&C メディア研究所

1. はじめに

コンピュータを利用する機会が増大するに伴い，人間のコミュニケーション特性を活かし，より自然な感覚での操作を実現することが可能なマルチモーダル・インタフェースの実現が望まれている．従来のマルチモーダル・インタフェース[1]は，時間軸を基準とした対応付けによりモダリティを統合している．しかし，操作として考えた場合には，時間的な基準は補足的に利用する必要はあるが，あくまで各モダリティの対話操作における意味を基準として対応付けを行うべきであると考えている．また，従来の研究は主に仮想世界を対象としたものであるが，実世界へ適応することがより重要であると考えている．

これに対して我々は，対象とする世界の知識記述や知識を利用したモダリティの意味解釈により，視線・指さしなどから「空間での注目情報」と，音声から「発話コンテキストでの注目情報」とを検出し，ユーザの注目情報をベースとして統合する枠組みを構築した．この枠組みを利用して，ユーザの指さしによる直接指示と自然言語での発話からユーザの注目情報を検出し，検出した注目情報を基に実世界の AV 機器を対話的に制御するシステムを試作したので報告する．

2. 提案方式

提案方式の特長は，ユーザから入力される各モダリティを統合する場合に，時間を基準とした制御に加えて，対象とする実世界(もしくは仮想世界)を記述した知識に基づき，操作の意味的な正しさをより重要な判断基準として用いている点にある．

2. 1 対象世界の知識記述

ユーザから入力される各モダリティの認識結果から，ユーザが何に注目してどのような意図を持っているのかを推定する．そのためには，ユーザが操作対象とする世界の知識が必要となる．図 1 に処理の流れを示し，図 1 中の具体例を利用して対象世界の知識記述[2]に関して述べる．例えば，AV 機器の制御を実現するためには，実世界に配置されている AV 機器の名称，機能，空間中での配置情報などが必要となってくる．この例では，“TV” は空間領域 1 に配置されており，“POWER”，“CHANNEL” などの機能があると分かる．また，ビデオの

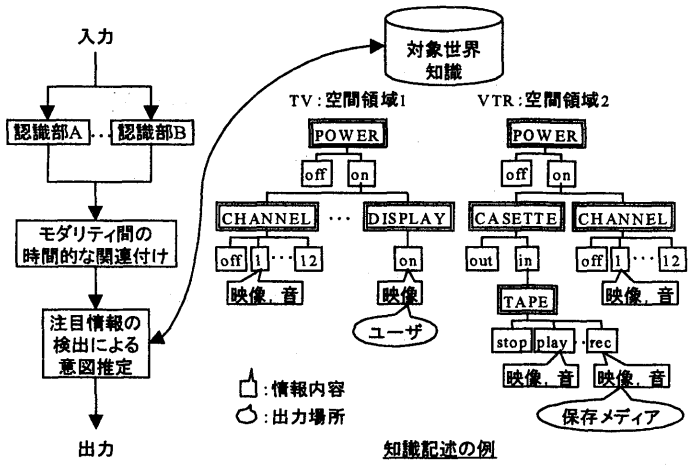

図 1：処理の流れ

「再生」とは「映像，音」という情報を出力するというような情報内容から機器間の接続関係を，出力場所から出力対象を特定できる．

2. 2 モダリティ間の時間的な関連付け

ユーザがジェスチャと音声による操作を行う場合，指さしなどのジェスチャと自然言語発話を同時に行うとは限らない．また，各モダリティを認識する部分での，認識処理の負荷による時間ずれも生じる．このような場合，本来同一の操作として扱うべきものであるのに，時間ずれが生じることになる．そこで，ある一定の時間間隔内において，モダリティ間に関連があるかどうかを検証することが必要となる．本方式では，各認識部から認識結果が出力される時間を基にして，お互いの信号が一定時間間隔内に出力されていれば時間的に関連性があると判断する．

また，各モダリティ毎に特性が異なるため，その差を吸収する必要もある．指さしによる領域指示では，単一操作で複数領域を指示した場合にも，各領域の指示は別の認識結果として出力される．一方，音声認識の結果は，発話の開始から終了までの一連の時系列がひとまとまりとして出力される．このために，従来の統合方式では，例えば，図 3 に示すように領域指示と発話コンテキスト内の指示詞との対応付けで不都合が生じる．また，指示詞は対象を指し示しているものなのか，文脈として現れるものなのかも分からない．従来方式では，文脈として現れる指示詞と領域指示との対応をとってしまうために，最終的

Multi-modal Interface with a method to detect User's Attention
Atsushi CHAEN*, Seiya OSADA**, Kazuo KUNIEDA*
* Human Media Res. Labs. NEC Corp.
** C&C Media Res. Labs. NEC Corp.

な意味として矛盾を生じてしまう場合がある．これに対して，本方式では，画像認識結果が短い時間間隔で連続して現れる場合には，単一操作での指示だと判断し，入力順序を保存した状態で複数信号を単一信号として処理するものとする．これにより，図 3 に示すように指示詞の数と領域指示数とが異なる場合にも，後段の処理で対応づけることが可能となる．また，文脈として現れた指示詞と領域指示とを 1 対 1 に対応づける必要は必ずしもなく，後段の処理で意味的なまとまりとして対応づけることが可能となる．

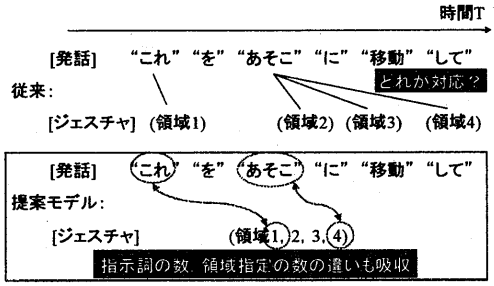

図 3：画像認識結果の扱い方による対応付けの違い

2. 3 注目情報の検出による意図推定

次に，ユーザの意図を推定し操作を実現するためには，モダリティ間の時間的な関連性の検証だけでなく，対象世界の知識を考慮した上で関連があるかどうかを判断する必要がある．これに関して，本方式では，図 1 に示すような知識記述を利用して「空間での注目情報」と「発話コンテキストでの注目情報」を検出し，時間的な関連性も含めた上で統合的に解釈を行っている．各注目情報とは次のようにして検出する．図 1 の知識記述を利用した検出例も併せて示す．

(1) 「空間での注目情報」：指さし指示による認識領域と知識記述との照合により検出される．領域指示結果が(領域 1)の場合には，“領域 1” という注目位置の他に “TV” という注目対象を検出できる．

(2) 「発話コンテキストでの注目情報」：発話コンテキストに含まれる名詞，動詞などと知識記述との照合により検出される．“つける” は(POWER,on)という知識と対応づけられており，知識記述のツリーから “TV” もしくは “VTR” を “つける” という注目情報を検出できる．

そして，時間的に関連がある場合には，これらの注目情報を相補的に補完することにより，操作が意味的に正しいものであるかどうかを判断する．しかし，時間的に関連があったとしても，意味的に正しくないと判断した場合には，ユーザとの対話により意図している操作を実現する．この例の場合は，“TV” を “つける” という実現可能な操作を導出できる．また，“TV” を指さし “再生((TAPE,play)に対応)” と発話した場合には，“TV” ではなく “VTR” に対する操作であると判断し，“VTR” で再生し “TV” に出力するという意味に解釈される．

2. 4 対話による注目情報の曖昧性解消

本方式ではシステム状態に応じてユーザへの問い合わせを生成し，ユーザとの対話により曖昧性の残った注目情報を特定することが可能である．例えば，指示語なしに “つけて” と発話するなどした場合，システムは知識記述から “つける” という機能を持っているものが複数あるかどうかを判断する．複数ある場合には，ユーザへ “どの機器をつけますか？” という問い合わせを生成する．この問い合わせへの応答として，システムは指さし，音声の両方に決定権を与える．また，“つける” という機能を持っているものに認識対象を限定するなどの制御が可能である．これにより，より自然でロバストな操作が実現できる．

3. 試作システム

設定したモダリティは指さしと自然言語発話である．対象機器としてテレビとビデオを用意した．カメラ入力による指さし認識処理，ヘッドセットマイクによる音声認識処理ならびに一連の統合処理を 1 台の PC で処理し，RS232C 制御の赤外線インタフェースにより実機器制御を実現している．指さしに関しては，ユーザは指定位置に立ち，腕をまっすぐに伸ばした状態で対象を指さしてもらうこととした．指さしの認識処理としては一般的な背景差分を利用した．

具体的な操作例としては，

・テレビ，ビデオそれぞれに対する電源のオン･オフ
・テレビを見ながらビデオのチャンネルを変更する
・ビデオを指示して再生させる

などである．これらの操作例に対して，認識処理による遅延はほとんど感じられずレスポンスは良好であった．音声認識は話者特定で学習をしており比較的安定して認識できている．また，指さし指示は指示方法を限定していることもありロバストに検出できている．

4. まとめ

統合モデルの検証を行う場として，AV 機器制御をターゲットに設定したマルチモーダル･インタフェースの試作を行った．限定された条件下ではあるが，応答性や誤解釈の少なさといった点で良好な結果を得た．本方式では，注目情報，ユーザとシステムとの対話状態に応じて，認識部にフィードバックをかける枠組みを有しており，これを用いて頑強性の向上を図る予定である．また，他のアプリケーション場への適用の検討，試作，知識記述の動的な更新方法の検討なども進める．

参考文献

[1] 市川，“マルチモーダル・インタフェースの動向と課題”，システム/制御/情報学会誌，Vol.39，No.5，pp.233-240(1995)
[2] 長田他，“自然言語を用いて家庭機器操作を行う対話システム”，信学技報 SP98-73，pp.23-30(1998)