グラフィクスとCAD 72−10
(1994. 12. 16)

指示棒と音声を用いるコミュニケーション環境 CoSMoSの設計（音声と統合処理）

山之内毅 大橋健 松永敦 江島俊朗

九州工業大学 情報工学部

〒820 福岡県飯塚市川津 680−4

あらまし 指示棒と音声が使えるマルチモーダルヒューマンインターフェイス環境CoSMoSの音声処理部の検討を行なった。情報環境CoSMoSのシステムでは、指示棒の動きはTVカメラからの画像情報により抽出され、音声情報はLPCケプストラムに変換されHMMにより認識される。本稿では、このシステムの音声処理部と（音声と画像情報の）統合処理部の構成法について検討を行なう。指示棒では位置情報を、音声では操作命令を指定しやすいことを考慮して、CoSMoSシステムの命令操作体系を組み立てた。また、指示棒を補完的に用いることにより、不要語や雑音による音声判別の煩わしさを軽減できる可能性があること、さらに意味的に同じことを表す指示棒の動作を音声に付加することにより、操作命令語の認識率の向上が望めることなどを考察した。

和文キーワード マルチモーダルインターフェイス、音声認識、隠れマルコフモデル

A Design of CoSMoS which is the communication environment with speech and motion of stick

Takeshi YAMANOUCHI Takeshi OHASHI Atsushi MATSUNAGA and Toshiaki EJIMA

Faculity of Computer Science and Systems Engineering,
Kyushu Institute of Technology
680−4 Kawazu , Iizuka, Fukuoka, 820, Japan

Abstract

We have proposed a multi-modal human interface called CoSMoS (Communication with Speech and Motion of Stick), in which speech and/or the stick are used to communicate with computer. In this paper, we investigate a better command system used in CoSMoS, where the stick is regarded as pointing device and speech are used as a command to system. Furthermore, an integration of information from speech and information from a motion of the stick is shown to promote the ability of the CoSMoS system.

英文 key words multi-modal interface, speech recognition, hidden Markov model

1 はじめに

従来のヒューマンマシンインターフェイスは、キーボードやマウスと言った少々特殊な器具を用いて行なわれてきたため、コミュニケーションが円滑に行なわれない場合も少なくない。したがって誰にでも使える身振り手振りや音声が使えたら非常に嬉しい。コンピュータの処理能力の向上によって、それらの手段を扱う研究が現実味を帯びて来た。マルチモーダルインターフェイスがそれである。我々は最近、指示棒と音声でコミュニケーションを行なう環境CoSMoS (Communication with Speech and Motion of Stick) を提案している [1]。

マルチモーダルなインターフェイスではユーザーの意図を伝達するときに、同時に複数の手法を使って表現することができる。解釈するコンピュータ側としてはこのように冗長性のある方が、認識精度を上げることができありがたい。ユーザー側としては、一見複数手段を用いることは無駄で面倒なことに思えるが、例えば会話をするときに無意識に手振りを交えるように、複数用いた方がしっくり来ることも多い。またそれぞれの手段に表現の得意分野があるので、表現したい内容に合った手段を用いられる方が、使い勝手が良くなる。

CoSMoSの特徴は手振りを直接用いる代わりに指示棒を用いてコミュニケーションを行なうことである。このことにより、手振りの表現力をあまり落さずに、動きの認識処理の計算コストを低減できる。また、音声については完成された文章や文を用いずに、単語単位で発声された言葉からキーワードを見つけ解釈するワードスポッティング的手法を用いる。

本論文ではCoSMoSの概要を説明し、具体的なアプリケーションを通じて使い勝手の良いインターフェースを設計する。そのためにはそのような方策を講じれば良いか、音声処理を中心にして述べる。

2 CoSMoSの概要

CoSMoSのシステム外観を図1に示す。

指示棒の動き（ジェスチャ）はＴＶカメラで、音声はマイクでとらえる。できることならユーザーには何の拘束もない方が理想的であろう。しかしこれらの装備は、プレゼンテーションなどでも普通に用いられる程度のものなので、ユーザーは特に違和感なく操作を行なうことができる。またカメラに指示棒が映る限り、ユーザーは自由に動くことができる。この特徴は、CoSMoSは通常のコンピュータの使用だけでなく、ステージ上でのプレゼンテーションなどにも応用可能にしている。より広い範囲をカバーしたい場合や、指示棒の精度を上げるためには複数のカメラを用意する。上方や側方からも撮れるようにカメラを設置するとより高精度の指示棒認識が行なえる。

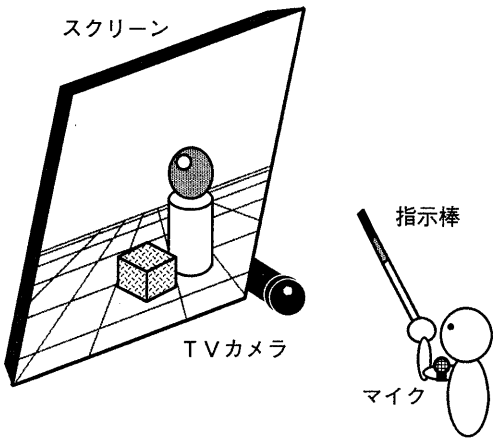

図 1: CoSMoSの外観

カメラとマイクより取り込んだデータはコンピュータに処理され、アプリケーションソフトの入力となる。これらの処理をリアルタイムに行なうには未だかなりの計算力を要求される。ここではその計算力の充足は近い将来実現されるものとして検討を行なう。

CoSMoSの構成は大きく4つに分かれる（図2）。

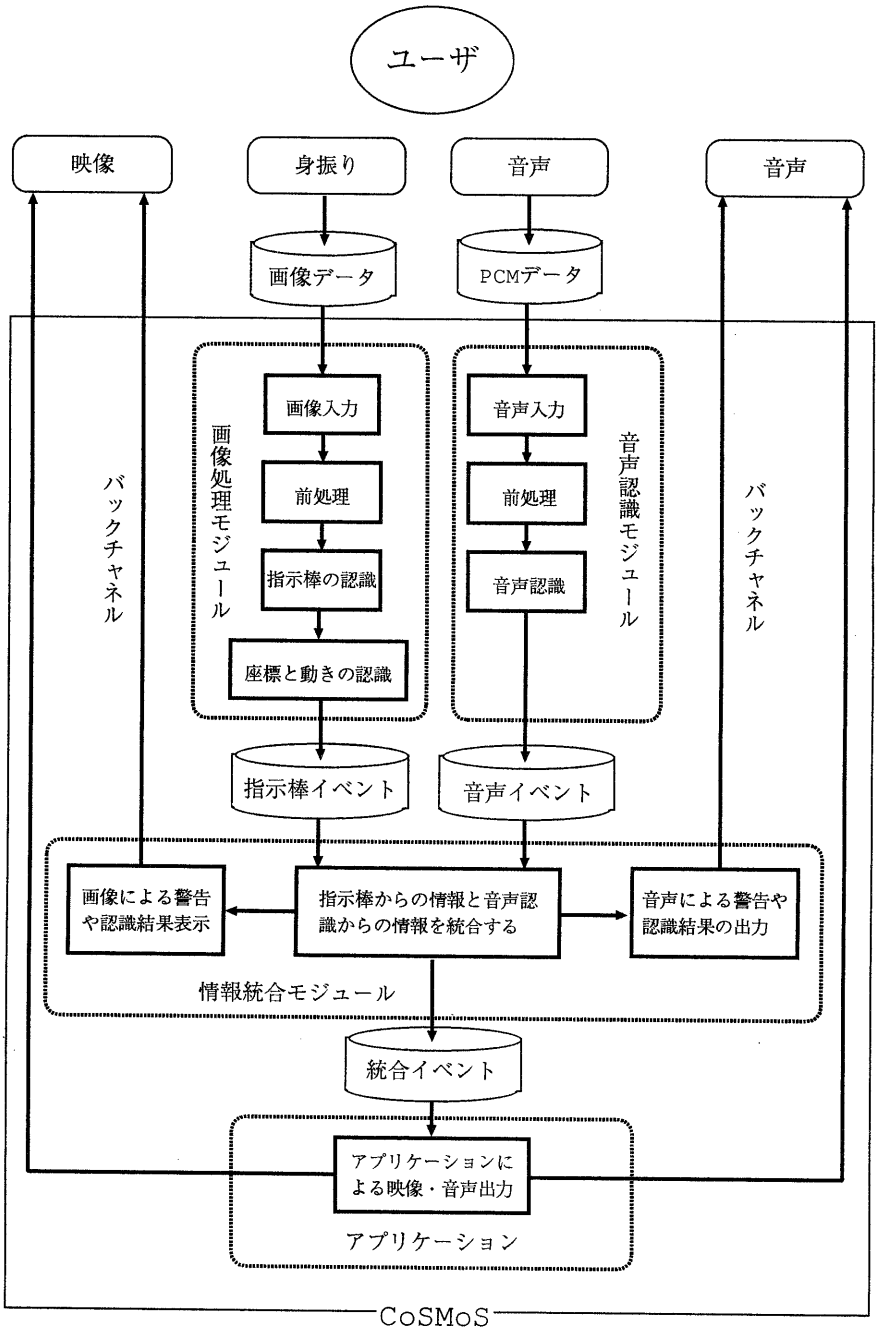

図 2: CoSMoS の構成

画像処理モジュール 取り込んだ画像から指示棒を抽出し、ユーザーが指している座標（3次元座標も可）や、一連の棒の動きから特定のジェスチャを認識する。

音声処理モジュール マイクより入力された音声を認識し、CoSMoSの命令やパラメータとする。

情報統合処理モジュール 画像、音声両モジュールからの情報を統合する。

アプリケーション（バックチャネル含む） 入力に対し処理を行ない、結果をユーザーに返す。

本論文では冒頭で述べたようにCoSMoSの音声処理から情報統合処理モジュールを中心に据えた検討を行なう。

3 アプリケーション「積木遊び」について

CoSMoSを具体的に論じるために、積木遊びを題材に取り上げる。このアプリケーションは、コンピュータにより設置された仮想空間内で、積木の生成、移動、消去などを行ない、目的の立体を構成するものである。積木遊びの操作体系を構築するために、まずこのアプリケーションにおいて、CoSMoSの二つの指定方法には得意／不得意とする指定事項があることを考慮する必要がある（表1）。

事項＼指定方法	音声	指示棒
動作	○ 「掴む」「動かす」など操作したいことを言う。口で簡単に指示できる。	△ 棒の動きのパターンで操作を指示する。音声に比べると直観的でなく、指示しにくい。
動作の修飾	○ 「速く」「少し」など口で言う。パラメータなどの数値も簡単に指定できる。	△ 操作と同じく棒の動きで示す。細かい指示はできない。
位置	△ 「〜の右」「少し奥」「3つ上」のように言う。細かい指示もできるが、面倒。	○ 棒でその座標を指すだけで良い。
形状	○ 「立方体」「円柱」など口で言う。簡単に指定できる。	○ 形状をを棒で描く。ユーザーが名称を知らない形状でも指定できる。
色	○ 「紫」「黄色」など口で言う。簡単に指定できる。	× 棒の動きで示すのはあまりに不便。

表 1: 音声と指示棒の特徴

命令やパラメータの指定は表1のように音声の方が得意であることから、音声中心に記述した。その一部を図3に示す。積木遊びとは異なるアプリケーションでは、指示棒中心の操作体系も考えられるであろう。

ユーザーの要求に柔軟に対処するため、操作指定と積木指定の順番は問わないようにした。また積木の指定は音声と指示棒のどちらでも可能である。例えば赤い立方体を左へ動かすには次のような手順がある。

手順例1）

- `grasp red cube`（赤い立方体がアクティブになる）
- `move`（積木を動かすモードになる。デフォルトは`location mode`）

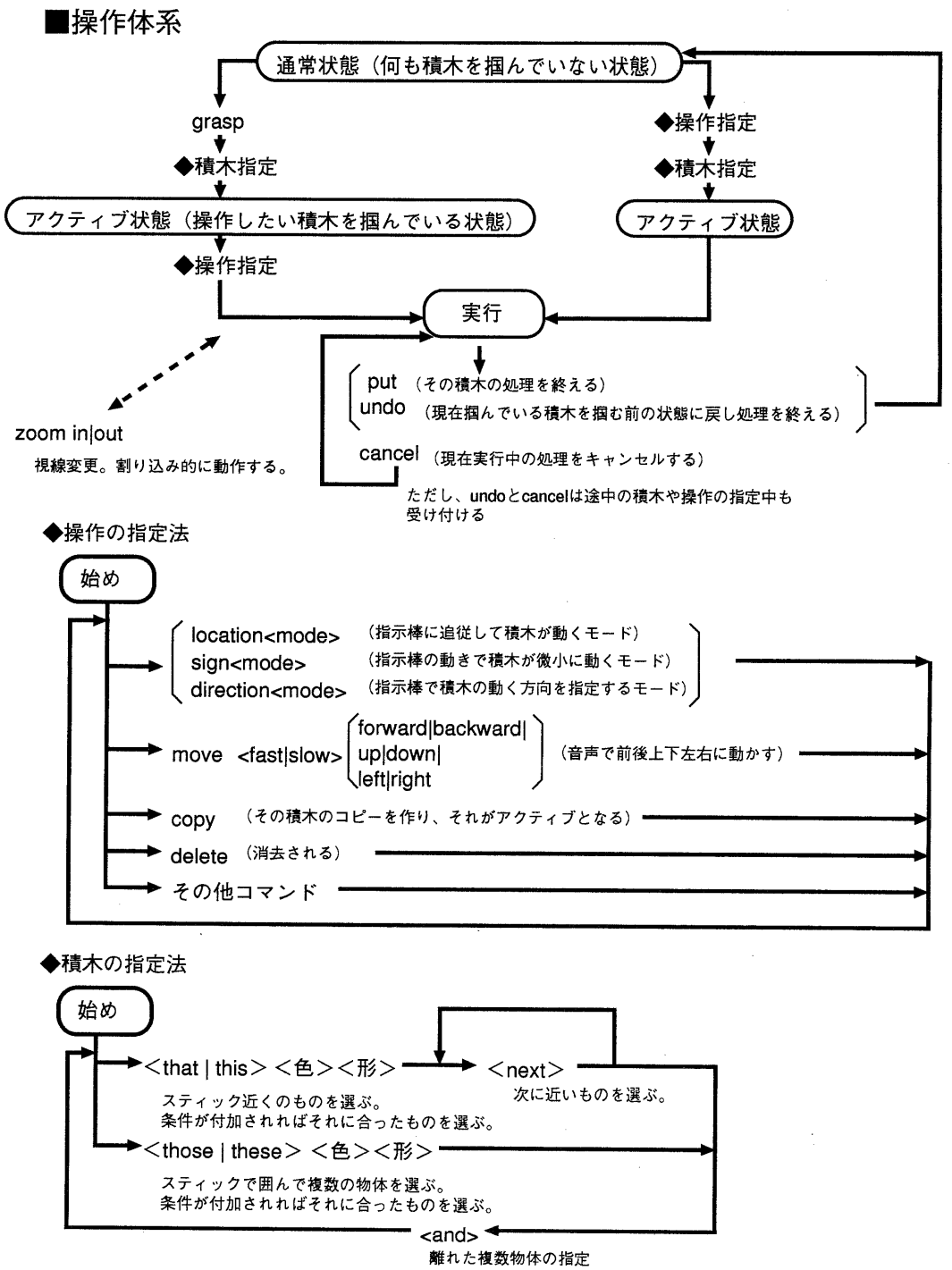

図 3: 積木遊びの操作体系

・ （棒で左へ動かす）
・put （目的の位置へ置いて終了）

手順例２）
・grasp （棒で指した赤い立方体がアクティブになる）
・move （積木を動かすモードになる）
・ （棒で左へ動かす）
・put （目的の位置へ置いて終了）

手順例３）
・move （積木を動かすモードになる）
・red cube （赤い立方体がアクティブになる）
・ （棒で左へ動かす）
・put （目的の位置へ置いて終了）

手順例４）
・move （積木を動かすモードになる）
・this （棒で指した赤い立方体がアクティブになる）
・ （棒で左へ動かす）
・put （目的の位置へ置いて終了）

このようにユーザーは自分の好みにに合わせて操作を行なうことができる。

4 音声認識手法について

音声認識はパワー情報を用いて発声を区切り、単語単位に行なう。これは認識処理が小さい計算コストで済むことと、１単語発声する毎にすぐに次の統合処理などを行ない、処理の遅延を少なくするためである。音声はマイクで取り込むが、その際にマイクが衣服と擦れたり、息が吹きかかったりするなどして、ノイズが発生する。よってこのノイズを除去するために前処理として、一定閾値以下のパワーが一定時間続いた場合は音量をすべて０にする処理を導入した。閾値を大きくとるほどノイズには強くなるが、逆にノイズでない部分もノイズに見なされてしまうため、ユーザーにはより大きくはっきりと発声してもらう必要がでてくる。

認識には隠れマルコフモデル（HMM）を使った。積木遊びで用いられる単語約５０語（図3に表記してある単語と数字）を登録語とし、各々についてモデルを生成しておく。認識時には前処理で区切られた単語の時系列データの、各モデルに対する尤度を計算する。

表2にHTK V1.5(HMM Toolkit version1.5)[2]を用いて、特定話者による発声で認識実験を行なった結果を示す。特徴はLPCケプストラム係数とパワーを使用した。HMMの状態数は10、次状態と次々状態への遷移を許すモデルである。学習では、登録語５０語を使って喋った連続音声中から、登録語を手で切りだしてそれをモデルの訓練に用いた。学習データにはのべ700語程度の登録単語が含まれている。またノイズのHMMを一つ作り、学習データの中に散在するノイズを使って訓練した。

認識は未知音声を前述の手法で切り出した、登録語200語、不要語50語で行なった。不要語は登録語でない英単語４０語衣服の擦れる音やマイクを叩いたりして発生させたノイズ１０語である。

認識器にはノイズのHMMを使わない普通の場合と、それを用いた場合の二つで行なった。表2の上の表は、ノイズのHMMを用いず、不要語の除去の方法として尤度が著しく下がった場合にのみリジェクトを行なったときの結果である。下の表はそれに加え、ノイズのHMMを使った場合の結果である。認識結果がノイズになったときもリジェクトを行なう。

前者は認識率91.0%と、対話的なインターフェイスを構築するには充分である。しかし不要語

ノイズのHMMなし

	認識率	誤認識率	リジェクト率
登録語	91.0%	7.5%	1.5%
不要語		80.0%	20.0%

ノイズのHMMあり

	認識率	誤認識率	リジェクト率
登録語	81.5%	4.5%	14.0%
不要語		24.0%	76.0%

表 2: 単語認識の結果

のリジェクト率は低く、不要語に対しても何らかの結果を返してしまい、大きな問題となる。後者は認識率が 81.5% と低くなってしまうが、不要語の 76% をリジェクトできている。リジェクトをした場合は入力を繰り返すだけでよいが、誤認識をした場合は間違った処理を行なってしまう。

現時点では、どちらの手法を採るかはケースバイケースである。周囲が騒がしく不要語やノイズが入りやすい環境であればノイズの HMM ありで、部屋に自分一人で使うのであれば認識率の高いノイズの HMM なしで使うのが良いだろう。

5 使い勝手と認識率向上のための方策

使い勝手の向上のために、ユーザーの発声した命令が完全でなくても、システムの方でできるだけ補完をした方が望ましい。また発声されるであろう単語の予測がつくことは、認識率の向上や計算コストの低減につながる。

5.1 各単語へのスコア計算

各単語に以下のような方法でスコアを算出し、音声認識結果の尤度に乗算または加算して、最終的な認識結果とする。

文法制約 積木のアプリケーションでは操作体系はかなり柔軟であるが、操作系から外れた単語は出現する。このような単語は操作系の文法から逸脱しているとして、0 に近いスコアを与える。実際にユーザーが間違って発声した場合は、補完し切れない状況になるので、リジェクトされ、ユーザーに再度指示を求める。

過去の履歴をとった情報の補完 物体、動作、位置など CoSMoS アプリケーションの中で指定できる各々の事項別に履歴をとって、スコアを算出する。頻度の高いものや最近に出たものと高スコアにする。例えば、$1/e^t + 1$（t はその単語が出現してからの経過時間）のようにスコアを与えると、一度も使われていない 1 倍から、直前に使われた 2 倍の範囲で尤度が変化する。

スティック情報 座標や大きさ、範囲などの指定は指示棒の得意分野である。これらのパラメータが必要なときは、入力された単語よりもより高いスコアを、指示棒からの入力に与える。

以上の処理の結果、尤度が一定閾値以下の時は、入力または認識処理に誤りがあると見なし、入力を無効とする。

5.2 スティック制約

システム稼働中、常に音声認識が働いていると、CoSMoS以外のことにユーザーが注意を向けてそちらに発声をしたときに、システムが余計な動作をする危険性がある。そこで、カメラに指示棒が映っていないときはシステムをポーズ状態におく。ユーザーはCoSMoSと対話をしたくないときは、カメラの範囲外に指示棒を出せばよい。

逆にCoSMoSに命令を出したい時はカメラの範囲に指示棒を入れる、CoSMoSに注意を向ける動作が必要である。このことでユーザーの注意力が高まり、よりスムーズなシステムとの対話が行なわれるという副次的な効果も期待できる。また複数のCoSMoSを同時に使うことも可能である。操作したいCoSMoSのカメラの視野に指示棒を入れれば良い。

5.3 バックチャネルの工夫

操作をしていてもシステムからの反応がなかったり、遅かったりするとユーザーは不安になり、使い勝手が悪くなる。CoSMoSのバックチャネルはアプリケーションに多く依存するが、最低限として、スティックカーソルの表示と認識した単語の表示は常にやっておく必要がある。また、システムが操作ミスや誤認識で意図しない動作をした場合に備え、常に実行中の操作のキャンセル、実行後のアンドゥーが簡単にできなければならない。積木遊びの操作体系にはそのあたりも含めてある。これらのことも使い勝手の向上には重要な要素である。

6 まとめ

指示棒と音声を用いるヒューマンコミュニケーション環境CoSMoSの概要を説明し、音声処理部分と統合処理部分を中心に設計した。また、CoSMoSで用いる音声認識処理の不要語の除去方法などについて検証し、指示棒と補完し合うことでシステムの使い勝手を向上する方策を検討した。

謝辞

本研究を行なうにあたって御協力頂いた情報処理機構講座の皆様に感謝致します。また本研究の一部は文部省科学研究費一般研究C（課題番号06808039）による。

参考文献

[1] 大橋, 山之内, 松永, 江島, 指示棒と音声が使えるコミュニケーション環境CoSMoSの提案, 第2回インタラクティブシステムとソフトウェアに関するワークショップ講演論文集, 1994

[2] S.J.Young,P.C.Woodland,W.J.Byrne, HTK: Hidden Markov Model Toolkit V1.5 Referance Manual,User Manual, Cambridge University Engineering Department Speech Group and Entropic Reserch Laboratories Inc., 1993

[3] Richard Bolt, Put-That-There: Voice and Gesture at the Graphics Interface, Computer Graphics.14[3].pp.262-270, 1980