1C1-04

# 確率的推論を利用したマルチモーダル対話制御

Controlling Multi-modal Dialog using Probabilistic Reasoning

麻生英樹[*1] Hideki ASOH　小玉智志[*2] Satoshi KODAMA　アブデラジズ・キアット[*3] Abdelaziz KHIAT　松本泰明[*4] Yasuaki MATSUMOTO
本村陽一[*1] Youichi MOTOMURA　原 功[*1] Isao HARA　浅野太[*1] Futoshi ASANO　新田恒雄[*2] Tsuneo NITTA　小笠原司[*3] Tsukasa OGASAWARA
柿倉正義[*4] Masayoshi KAKIKURA

[*1]産業技術総合研究所 National Inst. of Advanced Industrial Sci. and Tech., AIST
[*2]豊橋技術科学大学 Toyohashi Univ. of Tech.
[*3]奈良先端科学技術大学院大学 Nara Inst. of Sci. and Tech., NAIST
[*4]東京電機大学 Tokyo Denki Univ.

A method of estimating user's intention in multi-modal dialog using probabilistic reasoning is proposed. An architecture of dynamic Bayesian network is designed and applied to a simple data base retreaval task. An implementated multi-modal dialog system based on the idea is also described.

## 1. はじめに

マルチモーダル対話システムには，時々刻々とパターン情報から言語情報に至る様々な入力が与えられ，システムはそれらにリアルタイムに対応して適切な応答をすることが求められる．マルチモーダル対話システムが扱う情報は，その表層的表現の多様性に由来する不確実性に満ちている．カメラからの画像入力，マイクからの音声入力のいずれについても，入力情報の表層的なバリエーションは膨大であり，また，多くの情報が混在している．システムは，時系列である入力パターン情報を分節・認識・解釈して，そこから適切な意味情報や対話の状況に関する情報を抽出し，必要な問題解決を行って，出力を生成する．

こうした過程は，われわれが日常的に生きて，いろいろな出来事に対処している過程と同じであり，特別なものではない．逆に言えば，対話の制御は，より一般的な，知的システムによる不完全情報下での問題解決と同じ 枠組みで考察されるべきである [Paek 00]．対話システムにおいて特徴的なことがあるとすれば，システムの環境にあたるものの主要部分を対話相手である人間，すなわち最も複雑で不安定なもの，が占めるということと，人間との間で交わされる情報の主要部分が言語情報，すなわち，複雑な構造を持つ情報，であることだろう．その結果として，対話過程は，たとえば，地形が刻々と変化するフィールドでごく限られた感覚と言語記号を使ってナビゲーションするような不確実で不安定な過程になることが多い．

現在使われている多くの対話システムでは，問題を簡単にするために，不確実性をあらかじめ減少させている．すなわち，タスクを限定し，タスクを達成するためのシナリオを設定し，その狭いシナリオの範囲で対話を進行させる．しかしながら，そのような制約の強い対話はユーザにとって使いやすいものでないし，また，制約を破綻させる環境中の雑音に弱い．こうした制約の多さは，マルチモーダル対話システムが，ニーズの高さにもかかわらず十分に普及していない原因の１つであり，不確実性を柔軟に扱える対話システムの実現が求められている．

連絡先: 麻生英樹, 産業技術総合研究所情報処理研究部門, 〒305-8568 つくば市梅園 1-1-1 中央第 2, h.asoh@aist.go.jp

統計的なモデルにもとづく確率的推論は，環境の複雑さや不確実性に対処するためのツールである．確率的推論は，音声認識や画像認識などのパターン情報処理を中心として用いられ，その有効性が示されてきたが，近年，計算機パワーの増大にも支えられ，情報の多様性・複雑性に由来する不確実性の高い状況下での問題解決の汎用的なツールとしてより多様な問題へと応用範囲を広げつつある．

本稿では，対話中のユーザの意図の推定に確率的推論を利用する方法について述べる．さらに，その方法を簡単なデータベース検索対話に適用した結果と，現在構築中の対話システムについて報告する．

## 2. ユーザの意図の推定

対話システムの問題解決過程において，ユーザの意図や知識状態を推定することは重要な要素機能である [加藤 99]．特に，情報検索タスクや案内タスク，予約タスクなど多くのタスクにおいて，ユーザとシステムとの対話の大半は，ユーザの意図，すなわちサービス要求，の推定に費やされる．

対話中のユーザの意図の推移を，ユーザとのやりとりを通じて推測するためには，ユーザモデル，すなわちシステムからユーザへの応答とユーザの意図とから，ユーザの反応が生成される過程をモデル化したものを用意して，そのモデルを使って逆問題を解けばよい．ここでは，ユーザモデルの表現にダイナミックベイジアンネットワーク (dynamic Bayesian network: DBN)（動的信念ネットワーク dynamic belief network とも言う [Dean 89]）を用いる．DBN は相互に依存関係にある複数時系列の一般的なモデルで，ベイジアンネットを時間方向にコピーして展開した構造を持つ．

図１にユーザモデルの最も巨視的なレベルでの構造を示した．隠れ変数 $S_t$ は時刻 $t$ のユーザの意図を表現する（ユーザの意図は対話中に変化することがある）．ユーザにとっての入力 $I_t$ は対話システムからユーザへの応答であり，ユーザの出力 $O_t$ は，対話システムにとってのユーザの応答である．このようなモデルを用いれば，ユーザの意図を推定する問題は，観測可能な時系列 $I_1, ..., I_t$ および $O_1, ..., O_t$ から隠れ状態であ

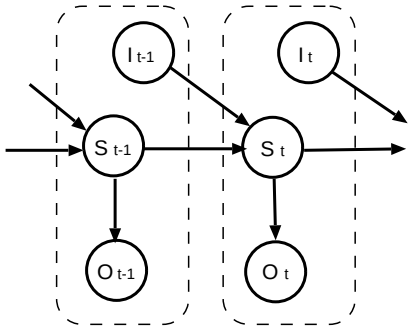


図 1: ユーザの意図推定のための確率ネットワーク

る $S_1, ..., S_t$ を推定する問題に帰着される．

このモデルを使うことで，新しい観測値が $I_t, O_t$ にセットされるたびに，それらを過去の情報と統合し，過去に遡って最も尤もらしい隠れ変数の状態遷移系列 $s_1, ..., s_t$ の推定しなおしたり，将来におけるユーザからの入力の予測を行ったりすることができる．

## 3. データベース検索タスクへの適用

### 3.1 対話の仕様

前節で示した計算論レベルの抽象的ユーザモデルは，巨視的な変数間のほとんど自明な依存構造を示しただけであり，実際のシステムに適用してゆくためには，$S_t, I_t, O_t$ を詳細化，具体化するとともに，その間の条件付き確率（の計算方式）を定める必要がある．

以下では，詳細化のプロセスを示すための例題として簡単なデータベース検索対話を取り上げる．データベース検索対話では，ユーザの意図は各種のデータベース操作コマンドと考えられるが，以下では，ユーザからの音声入力はすべて検索における指定項目の追加，SQL 文で言えば SELECT 文に対応するものとする．すなわち，1 つの対話中において，ユーザは検索要求を持ち，個々の発話では，その検索要求の一部を入力してゆくとする．たとえば，DARPA の ATIS プロジェクトにおける航空機スケジュールの検索タスク [DARPA 95] を例に取ると，ユーザは「金曜日の午前中のボストン行きの United のフライト」というような検索要求を持っており，その要求の一部を「金曜日のボストン行きのフライト」「午前中のフライト」といった形で入力してゆく．ユーザは，対話の途中で翻意して検索要求を変更したり，対話を途中で止めてシステムから離れてしまう可能性もある．

システムへの入力はユーザからの音声入力，ユーザの在・不在をモニタリングするためのカメラからの画像入力とする．システムからユーザへの応答出力は，ユーザからの検索要求に対する検索結果の視覚的な表示および検索要求の入力を促す音声とする．

### 3.2 ユーザモデルの詳細設計

時刻 $t$ は離散値を取るとし，なんらかのイベントが発生したとき，すなわち，視覚系からユーザの在・不在の変化が生じたときや，聴覚系から音声入力が生じたときに 1 つ進むとする．

前節のような対話の設定では，推定すべきユーザの意図 $S_t$ は，ある時点 $t$ でのユーザの検索要求および対話継続意志の有無を値として取る．簡単のためにこれらの変数間の関係は無視することとし，これらを 2 つのノード $Ss_t$ および $Sc_t$ によって表現する．観測変数 $O_t$ は，ユーザからの音声入力発話信号 $Oa_t$ およびユーザの在・不在をモニタリングするためのカメラからの入力画像 $Ov_t$ となる．また，$I_t$ は，システムが提示する検索結果 $Is_t$，入力を促す音声 $Ip_t$ の二つである．

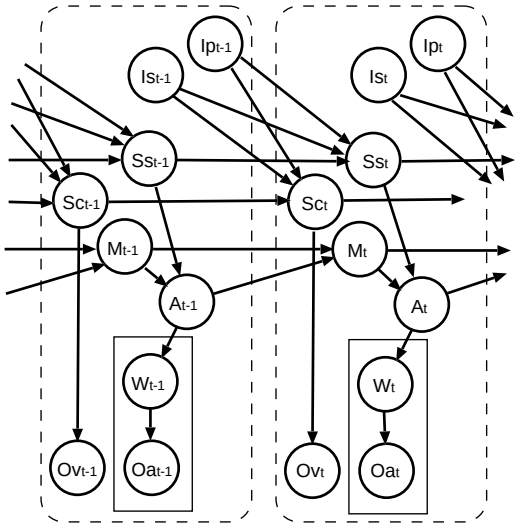


図 2: データベース検索タスク用の詳細な確率ネットワーク

ユーザの検索要求 $Ss_t$ や対話継続意志 $Sc_t$ は，システムからの情報である $Is_t, Ip_t$ に依存して確率的に遷移すると仮定する．

ユーザの意図が行動に変わるまでの過程は大変複雑であるが，ここでは，ユーザは現在の検索項目の中でどの項目を既にシステムに伝えたかを記憶しておき，その情報と現在の検索要求およびシステムからの情報を考慮して，次にどの項目を発話するかを決めるとする．そして，発話した結果として，これまでに伝えた項目の状態が変化する．これらを確率ネットワークによって表すために，既発話項目 $M_t$ および発話項目 $A_t$ という変数を導入する．$M_t$ の値は，$M_{t-1}$ および $A_{t-1}$ に従ってほとんど決定論的に推移するが，確率的な遷移にすることにより，対話の途中で記憶が混乱する効果をモデル化することもできる．

発話意図が決まると，確率的に発話文 $W_t$ が生成され，さらに確率的に音声信号 $Oa_t$ が生成される．一方，$Ov_t$ は，対話継続意図 $Sc_t$ に直接的に依存する．以上の変数の間の関係を示した DBN を図 2 に示した．図中の□で囲まれた部分が，通常の音声認識に対応する部分であり，このネットワークは，音声認識をマルチモーダルにし，意図レベルにまで自然に拡張したものになっていることがわかる．

## 4. マルチモーダル対話システムの実装

上記のユーザモデルを用いて，インタラクティブロボット用アプリケーション開発支援環境 [原 02] 上にデータベース検索マルチモーダル対話システムを実装した．

### 4.1 システムの構成

システムは，図 3 のように，視覚モジュール，音声認識モジュール，ロボットモジュール，音声合成モジュール，確率推論モジュール，データベースモジュール，統合モジュールの 7 つから成る．それぞれのモジュールは別プロセスとして TCP

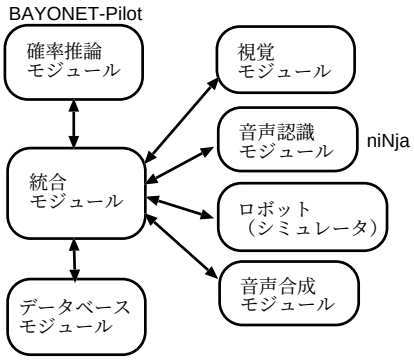


図 3: データベース検索対話システムの構成

で相互に通信しながら動作する．[*1]

視覚モジュールは，顔検出，視野内の動き（オプティカルフロー）計算などの機能を持ち，統合モジュールからのイベント検出要求に従って顔発見などのイベントをモニタリングし，検出したら通知する．音声モジュールは，統合モジュールからの要求に従って，不特定話者日本語連続音声認識システム niNja[Itou 93] からの出力を利用して音声入力の認識結果の上位 $N$ 位までを尤度値付きで返すことができる．辞書および文法はタスクのために作成したものを使用する．確率推論モジュールには，ベイジアンネット構築システム BAYONET-Pilot[本村 02b] に推論機能を追加したものを用いている．このモジュールは TCP の通信経由で確率ネットの構成を行い，観測値をセットして，推論を行わせることができる．ロボットモジュールはネットワーク経由で実ロボットあるいは Java3D を用いた 3D CG のロボットのさまざまな動きを制御できる他，複雑な動作を XML 形式で保存し，再利用することができる．音声合成モジュールは，富士通製の音声合成ソフトウェアを使用しており，他のモジュールと同様に，ネットワーク経由で制御できるようになっている．統合モジュールはすべてのモジュールを TCP/IP 接続を通じて制御し，ユーザとの対話を実現する．

試験的な検索対象データベースとして，囲碁棋士のデータベース[*2]を用意した．囲碁棋士に関する情報がレコードとして登録されている．個々のレコードは名前，所属（日本棋院，関西棋院など），師匠の名前，段位，取得したタイトル（複数），などのフィールドを持つ．ユーザは「日本棋院所属で９段で名人と本因坊になったことがある棋士は？」というような検索要求を持っていて，それをいくつかの文章に分けて音声で入力する．

### 4.2 確率的推論の実装

確率的推論の実装において最も問題となるのは，状態数の爆発である．たとえば，データベースが $L$ 個の属性（フィールド）を持ち，それぞれの取り得る値が $K_i(i = 1, ..., L)$ 通りとすると，検索要求 $Ss_t$ の取り得る値は $\prod_{i=1}^{L} K_i$ 通りとなる．これはかなり大きな数となるため，素朴な方法で状態遷移の条件付き確率を管理し，リアルタイムな応答を実現することは難しい．また，音声信号や画像信号は莫大なバリエーションを持つため，$W_t$ と $Oa_t$ あるいは，$Sc_t$ と $Ov_t$ の間の条件付き確率とそれを用いた推論の実装には工夫が必要になる．

今回の実装では，検索要求 $Ss_t$ は対話中には変化しないと仮定した．これによって，多様な $Ss_t$ を表現する必要がなくなる．対話中のユーザの検索要求の変化を含めてゆくことは今後の課題である．

一方，音声信号や画像信号のバリエーションに対応するためには，既存の音声認識モジュールや顔検出モジュールを利用する方法を考えた．既に述べたように，音声認識モジュールは，条件付き確率 $P(Oa_t|W_t)$ を計算するためのツールと考えられる．従って，$Oa_t$ は常に値 1 を取るダミーノードとし，そのノードの条件付確率テーブル $P(Oa_t = 1|A_t = a)$ に，音声認識システムの上位 $N$ 位までの候補出力の中で，発話内容が $A_t = a$ にあてはまるものの尤度値（複数ある場合には平均）を設定して確率伝播を行わせることにより，全体として必要な推論を近似的に実行することができる．画像入力 $Ov_t$ についても同様に，生のデータから在・不在を識別する処理は顔検出システムに任せて，その結果を利用した．現在の顔画像の検出モジュールは，尤度情報を出力しないため，出力結果に確率的ゆらぎを加えて条件付確率テーブル $P(Ov_t|Sc_t)$ を作成して推論を行わせた．

その他の条件付き確率については，論理的に妥当とおもわれる確率分布をプログラムによって生成して用いた．この作業は，従来の対話制御において状態遷移規則を記述する部分に相当する．実際にシステムを運用してゆき，対話事例が蓄積されれば，そのデータを用いて条件付き確率を修正してゆくことも考えられるが，今回は行っていない．

### 4.3 システム全体の動作

システム全体は以下のように動作する．

1. 起動されると，統合モジュールは，視覚モジュールと音声モジュールにユーザー検出のためのコマンド（顔検出依頼および挨拶入力検出依頼）を送り，待機する．
2. システムの前にユーザが座り，視覚検出モジュールから顔検出イベントが通知されると，統合モジュールはロボットシミュレータに，挨拶動作を依頼するとともに，挨拶音声および検索入力を促す音声の出力を音声合成モジュールに依頼し，音声認識モジュールに認識依頼を出し，視覚モジュールにユーザ検出を依頼する
3. 音声認識モジュールが発話音声を検出して認識結果を統合モジュールに返すと，統合モジュールは，ユーザの存在を確認し，確率推論モジュールに確率ネットワークの新しいセグメントを作成させ，認識結果を使って条件付き確率を設定し，過去の観測値も利用しつつ確率伝播[*3]により隠れ変数の確率分布を計算する．
4. 計算結果が与える最も尤もらしい解釈に従って検索要求の要素となるキーワードをスポッティングによって抽出し，検索要求にまとめあげ，データベース検索を実行して結果を画面表示する．
5. ユーザが検索終了音声入力を入力するか，視覚モジュールがユーザの不在を検出するまで 3, 4 を繰り返す．

今回の実装では，音声による入力プロンプトは，入力のタイミングを示すだけにとどまっており，特定の項目の入力を促すものではない．

この動作により，文脈の情報を利用しながら，過去の音声入力を再解釈してユーザの意図を推定することができる．これによって，ある時点では，音声認識の誤りによって所属＝日本棋

*1 以下の囲碁データベース検索の例ではデータベースモジュールと統合モジュールを一体として実装している．

*2 図４のシステムを実装した小玉の趣味による．

*3 厳密には Viterbi 計算が必要．

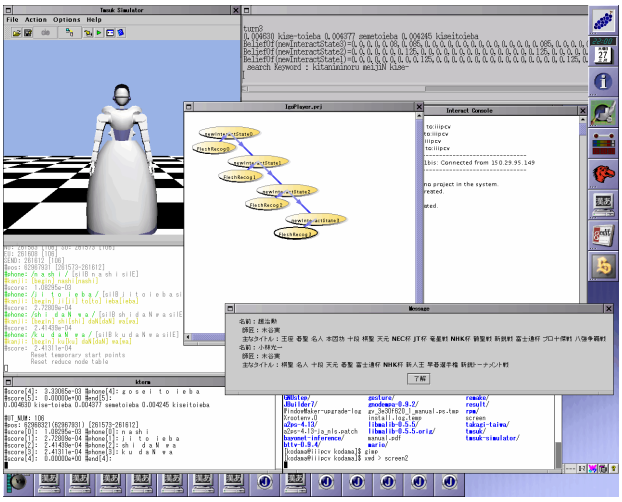

図 4: 囲碁棋士データベース検索対話システムの画面例

院，と解釈されていた音声入力が，後の時点での音声入力との文脈的な整合性の結果として，師匠＝木谷実，として正しく解釈しなおされる，といった効果が期待できる．

図４に実行中の画面のスナップショットを示す．左上にシミュレートされたロボットが表示され，中央には確率ネットワーク，右上にはそのネットワークの対話状況の確率分布が表示されている．

## 5. 関連研究

対話制御に確率モデルを導入する研究としては，[Akiba 94] が先駆的である．そこでは，渋谷の道案内の対話を対象として，ユーザの知識状態をベイジアンネットによってモデル化し，対話内容から，ユーザの知識に応じた適切な応答生成（知っていることは改めて言わない）を試みている．また，[乾 95] でも，ユーザの信念推定をベイジアンネットで行って，対話プラン生成や応答生成を行っている．[Singh 99, Levin 00, Roy 00] では，対話過程をマルコフ決定過程や部分観測マルコフ決定過程としてモデル化し，強化学習の枠組みを対話制御に適用することで，適切な対話戦略を学習させることが試みられている．[Paek 00] では，対話過程を様々な要因による不確実性に対処しつつ相互理解 (grounding) を維持しながら問題解決する行為と考える立場から，ベイジアンネットを用いて不確実性に対処するための対話制御アーキテクチャ *Quartet* を提案し，プレゼンテーション支援や受付案内に適用した例を示している．[本村 02a] では，プログラム仕様記述言語である UML (unified modeling language) を確率的に拡張した確率的タスクモデリングを提案し，インタラクティブシステムの制御に適用している．

このように，確率モデルを対話システムで用いる研究は，様々な試みが模索されている段階である．本研究は，ユーザモデルの構築という意味では [Akiba 94] および [乾 95] と類似しているが，用いられているユーザモデルは異なるものになっている．

## 6. まとめ

確率ネットワークを対話システムにおけるユーザの意図の推測に用いる方法を提案した．巨視的なネットワークの構造に基づいて，具体的なデータベース検索対話のための詳細なユーザモデルをデザインした．さらに，デザインに基づいて，実際に対話システムを構築した．今後，確率的推論を導入することの有効性の評価を行うとともに，さらに複雑なタスクに対する適用を試みてゆきたい．また，[岡本 02] では，確率モデルを用いてはいないが，対話システムの応答を Wizard of Oz 法による操作データから学習させてゆくことを試みている．こうした方法は，本稿で述べたような確率モデルの学習にも有効であると期待される．

**謝辞:** 本研究の一部は科研費 14208033 による．本研究で用いた対話システム用モジュールの一部は IPA 平成 12 年度および 13 年度未踏ソフトウェア創造事業による．本研究の一部は，小玉，キアット，松本が技術研修制度により産業技術総合研究所に滞在した間に行われた．関係者各位に感謝する．

## 参考文献

[Akiba 94] Akiba, T. and Tanaka, H.: A Bayeisan approach for user modeling in dialogue systems, *Proceedings of the International Conference on Computational Linguistics (COLING'94)*, pp.1212-1218 (1994).

[DARPA 95] *Proceedings of 1995 ARPA Spoken Language System Technology Workshop* (1995).

[Dean 89] Dean, T. and Kanazawa, K.: A model for reasoning about persistence and causation, *Computational Intelligence*, Vol.5, pp.142-150 (1989).

[原 02] 原, 本村, 麻生, 河村: インタラクティブ・ロボット基本ソフトウェアの開発, IPA ITX2002 発表論文集 (2002).

[乾 95] 乾健太郎: 自然言語生成における相互依存的制約の扱いに関する研究, 博士論文, 東京工業大学 (1995).

[Itou 93] Itou, K., Hayamizu, S., Tanaka, K., and Tanaka, H.: System design, data collection and evaluation of a speech dialogue system. *IEICE Transactions on Information and Systems*, Vol.E76-D, pp.121-127 (1993).

[加藤 95] 加藤恒昭: 対話システム, 田中穂積（監修）「自然言語処理－基礎と応用－」第 9 章, pp.281-381, （社）電子情報通信学会 (1995).

[Levin 00] Levin, E., Pieraccini, R., and Eckert, W.: A stochastic model of human-machine interaction for learning dialog strategies, *IEEE Transactions on Speech and Audio Processing*, Vol.8(1), pp.11-23 (2000).

[本村 02a] 本村, 佐藤: 確率的タスクモデリングによるインタラクティブシステムの制御と学習, 2002 年度人工知能学会全国大会（第 16 回）論文集 1B1-04 (2002).

[本村 02b] 本村陽一: ベイジアンネット構築ソフトウェア: BAYONET-PRO, ベイジアンネットセミナー BN2002 予稿集, pp.73-76 (2002).

[岡本 02] 岡本, 山中: Wizard of Oz 法を用いた対話型 Web エージェントの構築, 人工知能学会論文誌, Vol.17(3), pp.291-300 (2002).

[Paek 00] Paek, T. and Horvitz, E.: Conversation as action under uncertainty, *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI-2000)* (2000).

[Roy 00] Roy, N., Pineau, J., and Thrun, S.: Spoken dialog management for robots, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)* (2000).

[Singh 99] Singh, S., Kearns, M., Litman, D., and Walker, M.: Reinforcement learning for spoken dialog systems, *NIPS99* (1999).