Vol.2012-SLP-94 No.7
2012/12/20

対話ターン中のユーザ状態の推定に有用なモダリティの分析

千葉 祐弥[1,a)]　伊藤 仁[2]　伊藤 彰則[1]

概要：従来の音声対話システムは，ユーザが入力した発話を基準として処理を決定しているため，入力を待機している間にユーザの状態を推定することはできない．しかしながら，実環境下においてはユーザがシステムのプロンプトに戸惑ってしまい，入力をすることができない状況が度々起こる．このような場合，一定時間おきに同一内容のプロンプトを提示することが一般的であるが，この補助は入力内容を考えているユーザにとっては非常にわずらわしいものである．これらのユーザに対して適切な応答を行うためには，発話を行う前のユーザ状態を推定できる必要がある．以前行なっていた検討では，様々な影響を切り分けた分析を行わずに自動推定を試みていたため，どの情報がユーザの状態の推定に必要なのかが不明瞭であった．そこで，本稿ではあらためてデータの収集と被験者による評価実験を行い，より詳しい分析を行った．

1. はじめに

音声対話システムが柔軟な応対を行うためには，ユーザのモデル化によるユーザ状態の推定が必要である [1]．これまで，多くの研究が信念や嗜好 [2]，システムへの習熟度 [3], [4] といったユーザの内部状態に着目してきた．これらの研究は暗黙のうちに，対話システムがプロンプトを提示すれば，ユーザは常に入力を行うということを想定している．しかしながら，すべてのユーザがシステムを上手く使いこなせるとは限らない．例えば，システムのプロンプトの意図がわからなければ発話入力できず対話を放棄してしまうかもしれないし，入力が即答できるものでなければ入力内容を考える時間が必要である．

このような想定から，我々は少なくとも 2 つのユーザ状態が考慮されるべきだと考えた．一つは，ユーザがどんな入力を行うべきかわからない状態であり，もう一つはユーザがシステムのプロンプトへの入力を考えている状態である．ここでは，前者を State A，後者を State B とする．これらの状態は，従来の対話システムでは単に入力に時間がかかっているユーザとして同一に扱われてきた．本研究の目的は，上記の 2 状態に「円滑に対話できている状態」に相当する State C を加えた 3 つの状態を識別することである．

[1] 東北大学
Tohoku University
aoba 6-6-5, Aramaki, Aoba-ku, Sendai, Miyagi
[2] 東北工業大学
Tohoku Institute of Technology
Kasumicho 35-1, Yagiyama, Taihaku-ku, Sendai, Miyagi
[a)] yuya@spcom.ecei.tohoku.ac.jp

人間同士の対話においては，ある程度本研究が目指すような対話相手の状態推定が行われている．例えば，Feeling of Another's Knowing (FOAK) [5] と呼ばれる，対話相手が自分の発話に答えられそうかどうかを推量する能力などが当てはまる．FOAK に関しては，視覚的情報，音声情報との関連を調べた研究も存在する [6]．このような人間同士の対話のやり方を模倣することで音声対話システムの性能は向上できると考えられる．

以前までの報告では，システムのプロンプト発話を聞いてから，入力を行うまでのユーザの状態を対象としてきたが，元々の対話データの評価にはシステムの発話開始からユーザの入力発話終了までの全てを刺激として被験者に与えたため，どの程度ユーザの発話やシステムの質問に含まれる言語的情報が人間による推定結果に影響するのかが不明瞭なままであった．そこで，本稿ではより詳細な評価実験を行う．実験では異なる情報が含まれた 4 パターンの試料を作成し，ユーザの内部状態を三者択一で評価するよう評価者に求めた．ここでは，その実験結果を比較することで，対話に含まれる言語的情報と非言語的情報の重要性を分析する．

2. 対話データの収集

ユーザと対話システムのインタラクションを分析するため，Wizard of Oz (WOZ) 法に基づく対話収集実験を行った．WOZ 法は，操作された模擬的な対話システムとの対話を行ってもらう実験方法である．このとき，被験者はシステムが操作されていることを知らされないため，自然な状態での対話データを収集することができるというメリッ

トがある [7]．対話タスクにはシステムの質問に対して被験者が入力を行う，一問一答型のタスクを用意した．質問の内容は常識的な知識や事前に記憶してもらった数字列などを問うものである．

2.1 システム及びオペレータの動作

対話システムには被験者のシステムへの興味を持続させる目的で，エージェントが用意された．対話エージェントは顔のみが表示される非常に単純なものであり，オペレータが被験者の応答によって表情を制御する．オペレータは，1) 被験者にプロンプトを提示し，2) ユーザの入力に対して適切な応答を選択する．この際，被験者が装着しているピンマイクから入力音声を聞き，システムの応対を決定した．質問の提示順序は固定である．応答は被験者の入力が質問内容に対して妥当かどうかを基準に選択された．また，被験者には「わかりません」という入力も許可したが，実験の目的上，使用は最低限にとどめるよう求めた．質問のパターンは全部で 44 種類であり，質問に対して入力が行われない場合は，15 秒おきに同一内容のプロンプトを再提示した．

2.2 実験環境

図 1 に実験環境を示す．

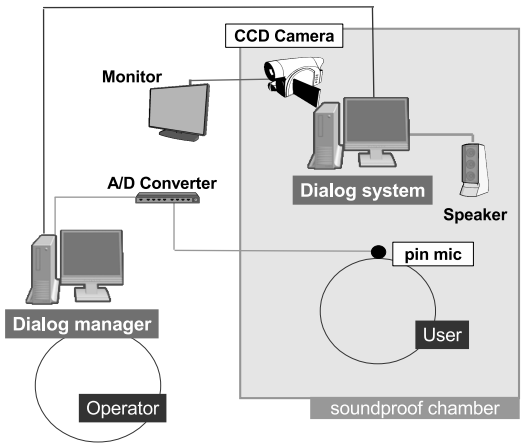


図 1 実験環境

実験は防音室内で行われた．防音室内の静音 PC のディスプレイには対話エージェントが表示され，システムの質問は防音室内の PC に接続したスピーカから再生される．オペレータは防音室外の PC からソケット通信によりエージェントに対話の指示を行う．被験者にはピンマイクを装着してもらい，音声を収録した．また，モニターの後部に CCD カメラを設置し正面から対話中の被験者の様子を撮影した．ビデオカメラの映像は収録と同時に，防音室外のモニターに映し出され，被験者の仕草を監視することができる．

2.3 収録条件

実験は男性 14 名，女性 2 名の計 16 名に対して行われた．ピンマイクの音声はサンプリング周波数 16 kHz，量子化ビット数 16 bit の WAVE ファイルとして保存した．動画像データは DV テープに保存し，実験後ビット深度 24 bit, 30 fps の AVI ファイルに変換した．

3. 評価実験

内部状態の評価にシステムの質問内容，被験者の仕草，音声のそれぞれがどのように関わっているのを調べるため，対話データの評価実験を行った．本稿で調査する課題は，

Q1: ユーザの入力内容はどの程度評価者の推定結果に影響するのか

Q2: 質問内容（プロンプト）はどの程度評価者の推定結果に影響するのか

Q3: 音声情報は評価者の推定結果に影響を与えるのか

の 3 つである．

3.1 セッションデータ

実験試料は，WOZ 法に従って集めた対話データを切り出したものである．一つのシステムの質問と一つの被験者の応答の組を一セッションとする．被験者の入力が行われずプロンプトを繰り返した場合は，システムの質問開始から再提示の直前までを一つのセッションとする．

3.2 評価実験用の試料

評価実験では上述した 3 つの課題に結論を出すため，一つのセッションに対して 4 パターンの試料を作成した．まず，各セッションを以下のように分割する．

(1) システムプロンプト区間の音声 A1 と動画像 V1
(2) プロンプト直後からユーザ入力直前までの区間の音声 A2 と動画像 V2
(3) ユーザ入力区間の音声 A3 と動画像 V3

実験試料はこれらの組み合わせで作成する．ここで用意したのは，

試料 A システムの質問開始からユーザの入力終了までを含んだデータ（V1，V2，V3，A1，A2，A3）

試料 B ユーザの入力区間を除いたデータ（V1，V2，A1，A2）

試料 C 試料 B のシステム発話区間をトーン信号でマスクしたデータ（V1，V2，A2）

試料 D 試料 C のトーン信号以外を無音としたデータ（V1，V2）

である．それぞれ，図 2，3，4，5 として示す．

評価実験の試料には，ビデオカメラによって収録した映像と音声を利用した．本研究では，自然な音声対話を実現するためには対話相手の内部状態の推測結果こそが重要であると考え，実際にどうだったか，すなわち真の対話相手

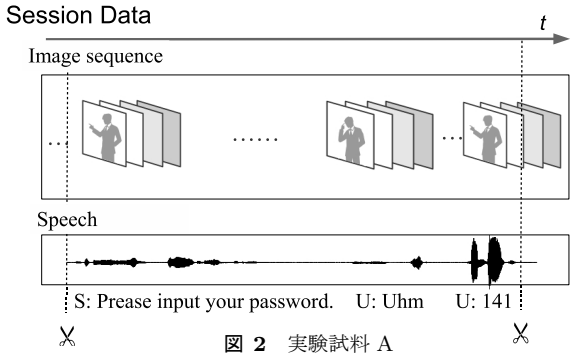


図 2 実験試料 A

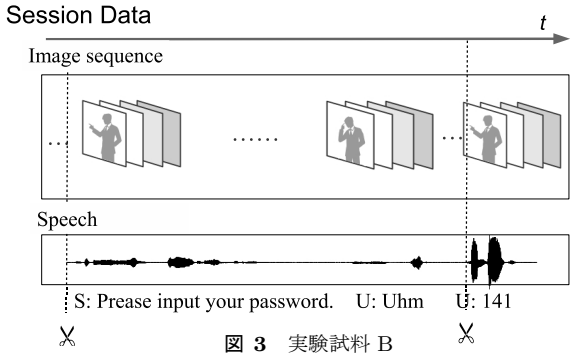


図 3 実験試料 B

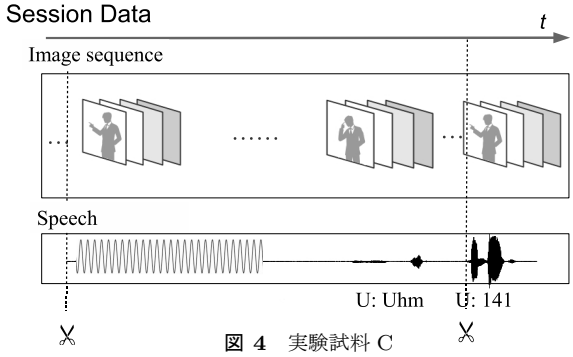


図 4 実験試料 C

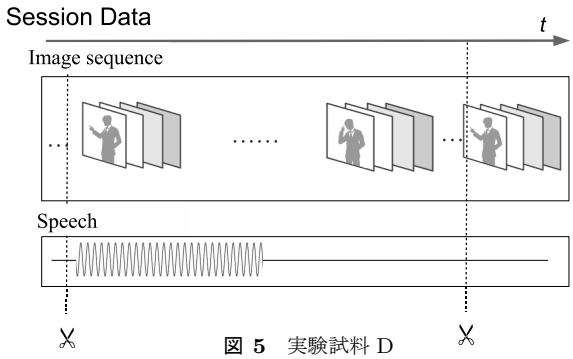


図 5 実験試料 D

の内部状態は問題としない．

3.3 対話データの選別

対話収集実験では，793 個のセッションデータを収集した．このうちの多くはシステムの質問終了後から即座にユーザの入力が起こっているデータである．このような対話では，ユーザの状態は State C と評価される傾向にあることがわかっている．そのため本稿では，評価実験に用いるデータとして，ある程度応答に時間が掛かっているものを選んだ．ここで，システム発話終了直後からユーザの入力が行われるまでの長さを latency と定義する．対象とするデータは latency が 5 秒以上となった 255 セッションとした．

3.4 実験手続き

評価は，対話収集実験に参加していない 18 名 (男性 13 名，女性 5 名) によって行われた．18 名を 4 つのグループに分け，それぞれ A 群，B 群，C 群，D 群とする．A 群の被験者は 3 名で，実験刺激として試料 A が与えられた．残りの被験者は 5 名ずつに分けられ，それぞれ試料 B，C，D が与えられた．評価者は各セッションを視聴した後，次のような設問に答えた．

問: 被験者はシステムの質問に対してどのような心境だったと思いますか．

評定 1) 尋ねられていることが不明瞭だと感じていた (State A)．

評定 2) 内容を把握し，質問に答えるために考えていた (State B)．

評定 3) 内容を把握し，迷わず質問の答えが決められた (State C)．

4. 評価実験の結果

4.1 A 群の結果の分析

ここでは，A 群の被験者 3 人の結果をまとめる．本稿では，一致率の指標として Cohen's κ を用いる．

表 1 に結果を示す．

表 1 Cohen's κ (A 群)

	E0	E1	E2
E0	—	0.46	0.53
E1	—	—	0.55
E2	—	—	—

表より，それぞれの評価者の評定は中程度の度合いで一致していることがわかる．また，三者間の一致係数を Fleiss's κ によって求めると，$\kappa = 0.51$ であった．これらの結果より，システムの質問開始からユーザの入力開始までの全ての情報が与えられた A 群の評価者の推定結果は比較的高い割合で一致していると言える．

4.2 A 群と B 群の比較

A 群と B 群の結果を比較し，被験者の応答内容がどの程度内部状態の推定結果に影響するのかを分析する．ここで，A 群の評価者の多数決の結果と B 群の評価者の多数決の結果の一致率は，$\kappa = 0.59$ となった．このことから，A 群と B 群の評価には大きな違いはなく，被験者の応答内容

による推定結果への影響は少ないと言える．ここで，質問に対する評定のばらつきの観点から結果を考察する．本稿では，評定のばらつきを平均情報量により数値化した．すなわち，

$$Ent_s = -\sum_j p_{sj}(G)\log_2 p_{sj}(G) \quad (1)$$

$$p_{sj}(G) = \frac{n_{sj}(G)}{\sum_{j'\in\{1,2,3\}} n_{sj'}(G)} \quad (2)$$

である．ここで，s はセッションの番号であり，$n_{sj}(G)$ は G 群の評価者がセッション s の試料に対して j と評価した回数である．平均情報量は評価が一致するほど低く，ばらつきが大きいほど高くなる．この平均情報量を用いて，A 群と B 群それぞれの中での評定のばらつきと A 群と B 群の間での一致・不一致の関係について調査した．

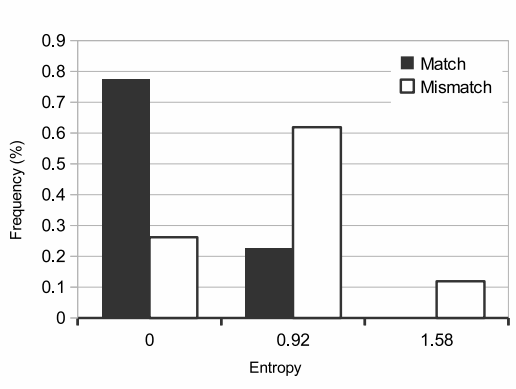


図 6 A 群の平均情報量の分布

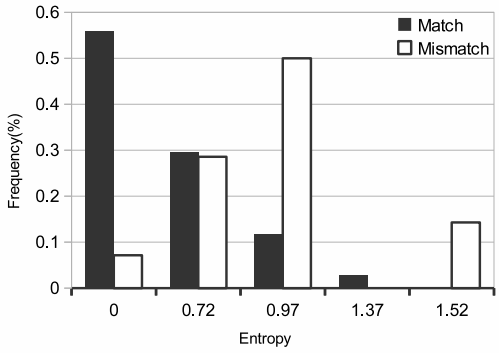


図 7 B 群の平均情報量の分布

図 6，7 は，A 群と B 群の平均情報量のヒストグラムを示している．黒色のビンは A 群と B 群で一致したものの平均情報量，白色のビンは一致しなかったものの平均情報量の分布を示している．

A 群と B 群の平均情報量は，いずれも一致したものと不一致だったものの平均値の間で有意差が得られている ($^*p < 1.0e-7$)．この結果から，A 群と B 群で推定結果が異なるものにはそもそも評価が難しいデータが多いということがわかる．実際に両群の平均情報量が小さいデータで評価が異なるものは少なかった．

以上から，ユーザの入力内容の内部状態推定への影響は少ないと言える．ただし，表 2 に示した通り B 群の評価者の個々の一致率自体には個人差がある．両群の平均情報量が小さく，評価が異なったものの例について述べる．このようなデータは質問内容の有無が評価に強い影響を与えたものであると考えられる．例えば，長い沈黙後に「わかりません」という入力を行ったユーザに関しては A 群は State A，B 群は State B と評価する傾向にあった．これは，このようなユーザが，B 群の評価者には入力を考えるために時間がかかっていたと判断されるが，A 群の評価者には考えたにも関わらずわからなかったので戸惑っていたのだと判断される傾向にあるからである．

表 2 Cohen's κ (B 群)

	E3	E4	E5	E6	E7
E3	—	0.34	0.46	0.35	0.56
E4	—	—	0.43	0.25	0.33
E5	—	—	—	0.23	0.55
E6	—	—	—	—	0.48
E7	—	—	—	—	—

4.3 質問が与えられたグループと C 群の比較

A，B 群と C 群を比較することでプロンプトの内容自体が推定結果に及ぼす影響を分析する．B 群と C 群の多数決の結果の一致率は Cohen's $\kappa = 0.30$ であり，あまり一致していない．

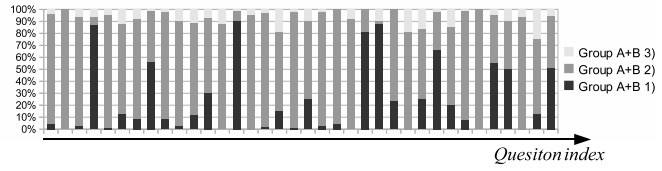
ここで，各質問に対してどのような評価がなされる傾向にあるかを調べる．この分析では，$n'_{qj}(G)$ を G 群の評価者が質問インデックス q の試料に対して j（$j \in \{1,2,3\}$）と評価した回数とし，

$$\mathrm{R}_q(G) = (r_{q1}(G), r_{q2}(G), r_{q3}(G)) \quad (3)$$

$$r_{qj}(G) = \frac{n'_{qj}(G)}{\sum_{j'\in\{1,2,3\}} n'_{qj'}(G)} \quad (4)$$

を計算する．$\mathrm{R}_q(G)$ は G 群の評価者の質問 q に対する評価の傾向を表すベクトルである．

図 8 は質問内容が与えられたグループ (A 群及び B 群) と C 群の質問に対する評価の割合を示している．図の横軸は質問番号を示している．対話収集実験では 44 個の質問を行ったが，8 つの質問については latency が 5 秒より大きくなるセッションが現れなかったため，36 個のビンが示されている．図 8 からは，質問内容を聞いたグループと聞いていないグループで評価の割合に似た傾向があることが見て取れる．そこで，それぞれの評価の割合について相関係数を計算した．結果を図 9 に示す．図 9 の点は A，B 群と C 群の $r_{q1}(G)$ の値に対応してプロットされており，36 点ある．このときの相関係数は 0.77 であり，高い相関が



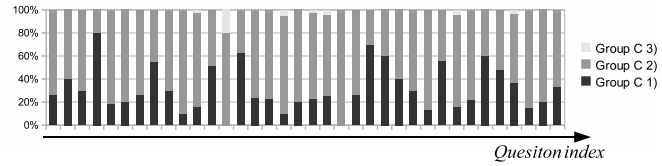


図 8 A,B 群と C 群の質問に対する評価の傾向

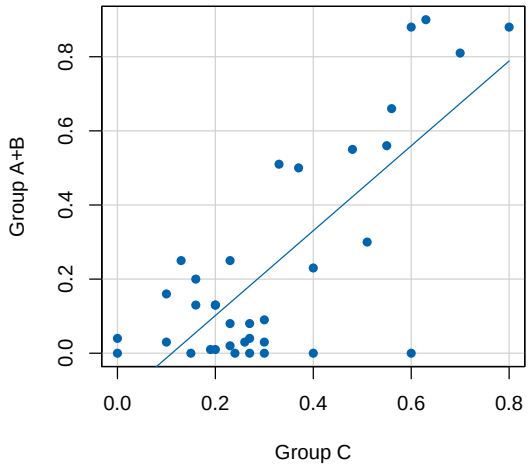


図 9 $r_{q1}(A \cup B)$ と $r_{q1}(C)$ による散布図

ある．

以上の結果から，質問内容を聞いたグループ (A 群及び B 群) と聞いていないグループ (C 群) の評価は多数決の結果の一致率は低いものの，全体的な評価の傾向では似た傾向を示していることがわかった．これは，質問の内容が被験者の状態推定に重要な役割を果たすが，言語情報に完全に依存するのではなく，入力までの非言語的な情報も影響を与えていることを示唆している．したがって，非言語情報のみから対話相手の内部状態を自動推定する試みには一定の効果が期待できる．

4.4 C 群と D 群の比較

以上までで，質問内容とユーザの入力内容の推定結果の影響を検討した．ここでは C 群と D 群の推定結果の比較を行うことで，視覚的情報と音声情報がどの程度対話相手の内部状態の推定に影響しているのかを分析する．ここで，C 群と D 群の多数決の評価の一致率は $\kappa = 0.41$ であり，

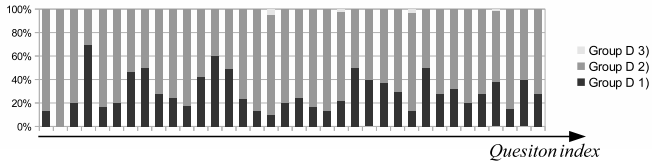


図 10 D 群の質問に対する評価の傾向

表 3 C 群と D 群の一致率 (Confusion Matrix)

		C 群				
		1	2	3	Nan	Total
D 群	1	28	17	0	0	45
	2	31	177	1	0	209
	3	0	0	0	0	0
	Nan	0	1	0	0	1
	Total	59	195	1	0	255

中程度の一致率である．これは，D 群の多数決の評価結果がやや評定 2 に偏っていることを反映した結果である．

図 10 より，D 群の評価結果の割合も，A+B 群，C 群とよく似た傾向を示すことがわかる．$r_{q1}(C)$ と $r_{q1}(D)$ の相関係数は 0.91 である．このことから，対話相手の内部状態の推定に関して，視覚的な情報は重要であると言える．

ここで，C 群と D 群の評価の一致率を Confusion Matrix として表 3 に示す．表中の Nan は，多数決の結果が同票でとなったセッションを示している．

C 群で 1 と評価され D 群で 2 と評価された 31 セッションと，C 群で 2 と評価され D 群で 1 と評価された 17 セッションでは，評定の変化に音声情報が大きな影響を与えていると考えられる．C 群で 1 と評価され，D 群で 2 と評価されたものに含まれる音声には，上昇長の短いフィラー，上昇長のつぶやき，含み笑いなどが多く観測される，この結果から，これらの音声は被験者の状態を State A と評価するための重要な特徴となると言える．

一方，C 群で 2 と評価され，D 群で 1 と評価されたセッションに含まれる音声には，長いフィラー，質問を復唱しているもの，長い独り言などが観測された．従って，上述した音声とともに，これらの音声特徴量も被験者の状態推定に重要であると考えられる．

また，評価結果が変化したものの中には，C 群においても明らかな被験者の音声が観察できないものがある．これらのデータは被験者が口唇は動かしているが実際には何もしゃべっていない，といったセッションである．このようなセッションは D 群では何かしゃべっているように見えるため State B と評価されるが，C 群では発声が確認できないため State A と評価される傾向があった．

以上のような結果から，入力発話前のユーザの内部状態の推定には視覚的な情報が大きく関与していることがわかった．しかしながら，C 群と D 群の評価結果の比較において示した通り，必ずしも視覚的情報だけで推定が行われ

表 4　各グループの RMS 距離

	A	B	C	D	Rand
A	0	0.24	0.32	0.33	0.44
B		0	0.3	0.31	0.43
C			0	0.25	0.35
D				0	0.34
Rand					0

ているわけではなく，部分的に音声情報が用いられていること，また，両者が同期的に関与する場合があることがわかった．

4.5　MDS による各グループ距離の視覚化

最後に，それぞれのグループの推定結果を多次元尺度構成法（MDS）によって二次元空間上へ写像した結果を示す．類似度の算出には RMS を用いた．すなわち，

$$d_{ij} = \sqrt{\frac{\sum_{s=1}^{N}(p_{s1}(i) - p_{s1}(j))^2}{N}} \qquad (5)$$

である．ここで，N はセッションの総数であり，$N = 255$ である．表 4 に，各グループ間の類似度を示す．比較のため，$p_{s1}(G = \text{“Rand”}) = 0.33$ としたランダム条件を加えた．表 4 に示したグループ間の距離を用いて MDS を行った結果を図 11 に示す． ここでは，A 群の評価結果が最も信頼出来る推定結果と考え，A 群の結果を基準とした考察を述べる．全てのグループの結果は，ランダム条件の結果よりも A 群に近く，与えられた情報がなんであれ，A 群の推定結果に近づくことがわかる．A 群と B 群の結果は特に近く配置され，前述したようにユーザの入力内容の推定結果への影響は少ないことを示している．同様に，C 群と D 群の距離から，音声情報を与えられた場合と視覚的情報のみを与えられた場合では，評価が似た傾向を示すことが読み取れる．一方で，A，B 群と C，D 群はやや離れた場所に配置されており，言語情報の影響の大きさが示された．

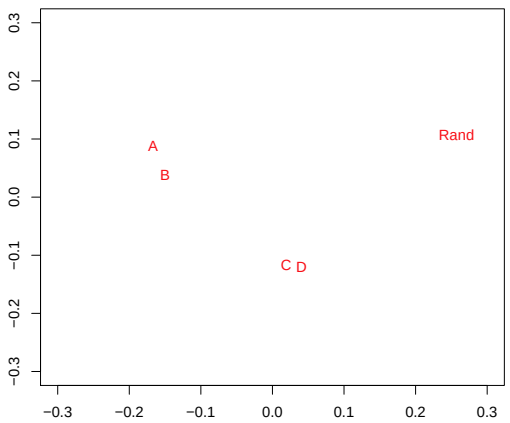


図 11　グループ間距離の可視化

しかしながら，A，B，C，D 群はほぼ直線上に並び，ランダム条件とは全く異なる評価傾向であることがわかる．これは，対話に関する情報が与えられると評価は一定の傾向でなされるようになることを示している．

5.　結論

本稿では，システムがプロンプトを提示してから入力を行うまでのユーザの内部状態に着目し分析を行った．以前までの検討では，ユーザの入力内容，プロンプト内容，視覚的情報と音声情報について詳細な検討をしないまま自動推定を試みていたので，そもそもの評価者の推定にどの程度各々の情報が関与しているのかが不明瞭であった．今回の分析により，1) ユーザの入力内容の評価への影響は少ない，2) プロンプトの内容は対話相手の内部状態推定に重要であるが，完全に依存するわけではなく，プロンプト提示後の対話相手の非言語的情報にも影響を受ける，3) 対話相手の内部状態推定には視覚的情報が大きく作用するが，一部音声情報が推定に用いられること，また，同期的に関与する場合があること，の 3 つを確かめた．

今後は，視覚的，音声的情報を同期的に扱うことでマルチモーダル対話システムのユーザの入力前の内部状態を推定する手法を検討し，実際にシステムへの実装を目指す．

謝辞　本論文は、総務省の「大規模災害時における移動通信ネットワーク動的制御技術の研究開発」（平成 23 年度一般会計補正予算（第 3 号））による委託を受けて実施した研究開発による成果である．

参考文献

[1] A. Kobsa. User modeling in dialog systems: Potentials and hazards. *AI&Society*, 4:214–231, 1990.
[2] A. N. Pargellis, H.-K. J. Kuo, and C.-H. Lee. An automatic dialogue generation platform for personalized dialogue applications. *Speech Communication*, 42:329–351, 2004.
[3] K. Jokinen and K. Kanto. User expertise modelling and adaptivity in a speech-based e-mail system. In *Proc. COLING*, 2004.
[4] F. de Rosis, N. Novielli, V. Carofiglio, A. Cavalluzzi, and B. De Carolis. User modeling and adaptation in health promotion dialogs with an animated character. *J. Biomedical Informatics*, 39:514–531, 2006.
[5] S. E. Brennan and M. Williams. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *J. Memory and Language*, 34(3):383–398, 1995.
[6] M. Swerts and E. Krahmer. Audiovisual prosody and feeling of knowing. *J. Memory and Language*, 53(1):81–94, 2005.
[7] A.Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. Desperately seeking emotions or: Actors, wizards, and human beings. In *SpeechEmotion*, pages 195–200, 2000.