音声言語情報処理 15−23
（1997. 2. 8）

対話の状態変化における統計的モデル化

森川 恵美　横山 真男　杉田 洋介　白井 克彦
早稲田大学理工学部
〒169 東京都新宿区大久保 3-4-1
morikawa@shirai.info.waseda.ac.jp

あらまし

近年，音声認識技術や音声言語処理技術の進歩により，高い音声認識性能が得られるようになってきた．しかし，それらの技術を駆使して構築されている現在の音声対話システムはタスク遂行の目的のために作られているため，ユーザの発話も機械的になってしまい，音声の利点が必ずしも生かされていないと思われる．本研究では，対話参加者がお互いに対等な立場をとることが可能であるタスクの対話データを利用して対話全体の中でのそれぞれの発話の役割を考慮して統計的に対話をモデル化し，対話の状態変化の様子を探ることを考える．統計的モデル化の手法として，現在音声認識技術として一般的に用いられている隠れマルコフモデル (HMM) を利用する．本稿では，まずモデル化に利用した対話コーパスと発話の分類について述べ，Ergodic HMM を利用した対話モデルを示す．そして情報量を利用したモデルの評価を行ない，モデルの応用例としてモデルを用いたタスクの識別とシステムへの応用例として次発話予測を試みる．

Spoken Dialogue Modeling Based on Statistical Approach

Emi Morikawa Masao Yokoyama Yosuke Sugita Katsuhiko Shirai
School of Science and Engineering, Waseda University
3-4-1 Okubo Shinjuku-ku Tokyo 169
morikawa@shirai.info.waseda.ac.jp

abstract

Recent progress of fundamental speech processing technologies such as speech signal processing, statistical modeling and language processing, has increased the performance of speech understanding. However, since existing spoken dialogue systems are made to achieve a task, the utterances of the user tend to be unnatural, restraining the effectiveness of speech interface. In our research, we analyzed dialogue corpora in which the subjects were on an equal standpoint, and modelized the dialogues based on the role of utterances, using the Hidden Markov Model(HMM). In this manuscript, we first described our dialogue corpora and the classification of utterances, and showed the spoken dialogue models made with Ergodic HMM. We evaluated our models by the calculation of entropy, and made a trial of task-identification and utterance prediction as examples of the application of our models.

1 はじめに

人間は音声を用いた対話によって人間同士のコミュニケーションを行っており，音声は人間にとってもっとも自然なコミュニケーションの手段であると考えられる．このことは，人間と計算機の間のコミュニケーションにおいても同様である．近年，音声認識技術や音声言語処理技術の進歩により，高い音声認識性能が得られるようになってきた．しかし，それらの技術を駆使して構築されている現在の音声対話システムはタスク遂行の目的のために作られているため，ユーザの発話も機械的になってしまい，音声の利点が必ずしも生かされていないと思われる．人間と自然に対話できるようなシステムを実現するためには，音響情報だけではなく，言語的知識やより高次な対話管理の知識を利用する必要がある．まず人間の音声を用いたコミュニケーションにおける振る舞い，またその背景に起こるさまざまな現象を分析することが必要であると考えられる．実際に人間同士の対話を分析してみると，タスクを与えた対話でもその発話の半分以上がタスク遂行には必ずしも必要でないもの (タスク共通発話) であることがわかる [1].

そこで本研究では，対話参加者がお互いに対等な立場をとることが可能であるタスクの対話データを利用して対話全体の中でのそれぞれの発話の役割を考慮して統計的に対話をモデル化し，対話の状態変化の様子を探ることを考える．統計的モデル化の手法としては，隠れマルコフモデル (HMM) を利用する．

本稿では，まず第 2章でモデル化に利用した対話コーパスと発話の分類について述べ，第3章でErgodic HMM を利用した対話モデルを示す．第4章に情報量を利用したモデルの評価を行ない，第5章にモデルの応用例としてモデルを用いたタスクの識別とシステムへの応用例として次発話予測を試みる．

2 対話データ

近年，多くの研究機関において様々な音声対話データの収集，分析が行なわれているが，それらの分析は 1 つのタスクの対話に注目したものやタスクの特徴の分析のために複数のタスクの対話で比較を行っているものが多い．従って，タスクに依存しない発話に注目して対話全体を分析している例は少ない．

本研究では，音声対話で現実に起こりうる多様な現象についてモデル化することを目標に，タスクに依存しない発話に注目して分析を行なう．

2.1 対話コーパス

分析に利用する対話コーパスは以下の 2 つのタスクによるものとする．

task1 クロスワードパズルタスク (8 対話使用)
2 人の話者がそれぞれクロスワードパズルの縦のヒントあるいは横のヒントを持ち，互いの情報を交換し合いながら 1 つのクロスワードパズルを完成させるものである．4 種類のパズルを使用する [2].

task2 スケジューリングタスク (CD-ROM[3] より 6 対話使用)
2 人の話者がそれぞれのスケジュール表を見ながら，会議などの日程を決めるものである．それぞれのスケジュール表は 1 カ月のうち 2 週間程度が埋まっている．被験者は教官の秘書であると仮定して会話を行なうものとする [4].

2.2 発話の分類

それぞれのタスク内での発話をみると，それらはタスクを遂行する上で必要となる情報を伝えようとしているものとタスク遂行には必ずしも必要でないと思われる情報を伝えているものがあることがわかる．例えばクロスワードパズルにおいては理論上，問題の場所と答，あるいはヒントのみが伝わればタスクは遂行できると考えられるが実際の対話にはその他の情報を伝達している発話が多く現れている．表1にタスク共通発話と各タスクのタスク依存発話の割合を示す．

表 1: 各タスクでのタスク依存発話の割合

	task 1	task 2	全体
タスク共通	912	222	1134
タスク依存	602	116	718
合計	1514	338	1852
タスク依存/合計	40%	34%	39%

2.3 タスク共通発話の分析

表1よりそれぞれの発話の中にはタスクに依存したものとそうでないものがあることが分かった．そこで，それぞれのタスクの発話に対して共通の分類を考え，ラベル付けを行なう．

タスクに依存した情報を伝えているものは1つのラベルとする．タスクに依存しない発話には対話自体を円滑に行なえるようにさせる様々な機能があると考え，タスクに依存しない発話を大分類することを試みる．

その結果，それぞれの発話に対して以下の12種類の発話ラベルを付加する．ここでは，1発話に対して複数のラベル付けを許し，また1単語に対しても複数のラベルを許した．表2に各発話ラベルとそれぞれの発話ラベルの出現数を，表3にそれぞれのタスクにおける1対話中の発話ラベル出現数の最大値，最小値，平均値を示す．

表 2: 発話ラベルと各タスクでの出現数

発話ラベル		説明・例
1	聞き返し	「えっ?」「もう1回言って」(認識失敗も含む)
2	呼びかけ	「それにしよう」(同意要求も含む)
3	疑問文	「わかる?」「いいですか?」
4	発話要求	「それで?」「で?」
5	同意肯定	「はい」「そうだね」
6	否定	「ちがいます」
7	返答	同意否定でない質問に対する答え
8	あいづち	「うん」「はい」
9	復唱	相手の発話の繰り返し
10	つなぎ語	「うーん」「えーと」
11	独り言	内容に影響しない発話
12	情報提示	タスクに依存した情報を伝えている発話

表 3: 1対話中の発話ラベル出現数

	task1	task2
最大	332	67
最小	105	46
平均	189.3	56.3

2.4 発話権の管理

第2.3節で1発話に対して複数のラベル付けを許し，また1単語に対しても複数のラベルを許してラベル付けを行なった．さらに，実際にそのラベルの発話の後，話者が交替するか否かを考慮すると各発話ラベルの出現数は表4のようになる．

表 4: 話者交替を考慮した発話ラベル出現数

発話ラベル		task1		task2	
		話者無	話者有	話者無	話者有
1	聞き返し	5	28	2	5
2	呼びかけ	28	103	30	22
3	疑問文	4	55	30	21
4	発話要求	0	2	0	0
5	同意肯定	65	71	6	31
6	否定	1	1	1	0
7	返答	19	31	7	28
8	あいづち	16	94	3	3
9	復唱	82	76	2	0
10	つなぎ語	141	24	29	1
11	独り言	27	39	1	0
12	情報提示	283	319	88	28
	合計	671	843	199	139

2.5 対話構造

クロスワードパズルタスク対話の特徴として1つの答を出すまでを1つの単位として対話全体が構成されていることがわかっている[5]．またスケジューリングタスク対話においても日時場所などを，順に決定していく．このようにどちらのタスク対話においても，1対話中に複数の話題が転換していく．以下の分析にはそれぞれの対話における話題の転換点にも注目する．

表5に1対話中の話題数の，表6に1話題中の発話ラベル出現数のそれぞれ最大値，最小値，平均値を示す．図1にクロスワードパズルタスク対話の発話例の一部とラベルを示す．

表 5: 1対話中の話題数

	task1	task2
最大	38	8
最小	23	3
平均	28.6	4.8

表 6: 1話題中の発話ラベル出現数

	task1	task2
最大	35	39
最小	2	1
平均	6.6	11.7

A： えーと,	< つなぎ語 >
1番は	< 情報提示 > 話者交替
B： うん	< あいづち > 話者交替
A： ビートたけしのギャグっていうと	< 情報提示 >
コマネチかな	< 情報提示 > 話者交替
B： コマネチ	< 復唱 >
B： じゃあ,	< つなぎ語 >
1の縦	< 情報提示 >
塩味だけの牛肉の缶詰で,	< 情報提示 >
コンビーフ	< 情報提示 > 話者交替
A： コンビーフ	< 復唱 >
⋮	⋮

"["でくくられた部分が1つの話題

図 1: クロスワードパズルタスク対話の発話例

3 Ergodic HMM を用いた対話のモデル化

現在音声認識技術として一般的に用いられている隠れマルコフモデル (HMM) を利用して，対話全体の中でのそれぞれの発話の対話中での役割 (発話ラベル) を考慮して統計的に対話をモデル化する．

task1，task2 の対話を対話全体を考慮して学習した状態数 5 のモデルをそれぞれ図 2,3に示し，対話全体を考慮し，話者交替情報を利用して学習したモデルを図 4,5 に示す．

図 2〜5において矢印毎の始めの数字が各状態間の遷移確率を示し，"[ ]" 内が発話ラベル番号を，後ろの数字が出力確率を示す．

また，図 4,5における "[ ]" 内の "s" は次の発話から話者が交替することを示している．

なお図 2〜5には遷移確率，出力確率ともに *0.1* 以上 のもののみを表示している．

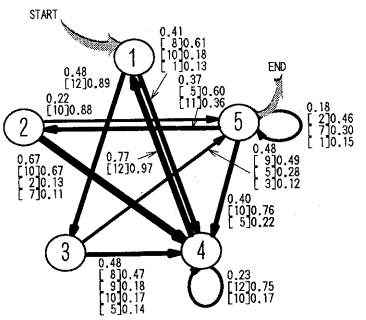

図 2: クロスワードパズルタスク (対話全体)

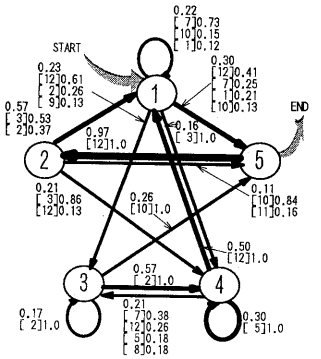

図 3: スケジューリングタスク (対話全体)

4 情報量を利用したモデルの評価

状態数 2〜7 の Ergodic HMM および発話ラベル遷移の k-gram の情報量を表 7 に示す．使用したモデルは第 3章で示したモデルと同様に開始状態と最終状態をそれぞれ固定にしている．

表 7: モデルのエントロピー

モデル	クロスワードパズルタスク				スケジューリングタスク			
	対話全体		話題毎		対話全体		話題毎	
	話者無	話者有	話者無	話者有	話者無	話者有	話者無	話者有
状態 2	2.20	2.65	2.53	3.15	2.25	2.93	2.16	2.19
状態 3	1.77	2.24	2.20	2.37	2.17	2.74	1.17	1.63
状態 4	1.68	1.79	1.90	1.36	1.66	2.44	1.45	1.27
状態 5	1.56	1.44	1.65	1.75	1.71	2.13	0.56	1.20
状態 6	1.10	1.51	1.46	2.04	1.48	1.76	1.46	1.10
状態 7	1.34	1.69	1.26	1.80	1.38	2.12	0.91	0.96
bigram	2.48	2.92	2.47	2.85	2.30	2.65	2.35	2.64
trigram	2.12	2.08	2.09	1.95	1.88	1.67	1.85	1.64
4-gram	1.51	1.91	1.34	1.07	1.51	0.71	1.17	0.73

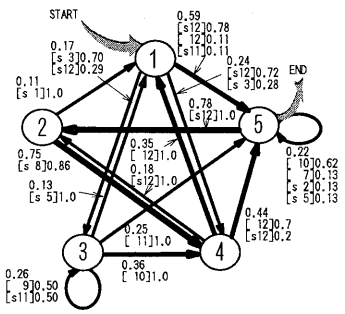

図 4: クロスワードパズルタスク (対話全体, 話者交替利用)

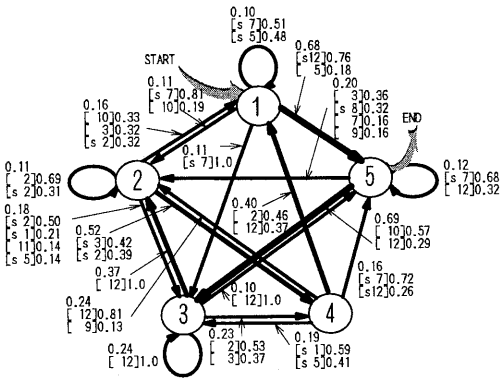

図 5: スケジューリングタスク (対話全体, 話者交替利用)

表中の "話者有", "話者無" は話者交替情報を利用したか否かを表し, "話題毎" とは 1 つの話題についての対話を 1 つの発話ラベル遷移列を使用したことを表している. 表 7 より, 一般的な傾向として, Ergodic HMM の状態数が増えるに従いエントロピーが小さくなり同じ状態数では話者交替情報を利用したものの方がエントロピーが大きくなっている. またスケジューリングタスク対話では対話全体を考慮したものよりも話題毎に発話ラベル遷移列を使用したものの方がエントロピーが大きくなっているが, スケジューリングタスク対話では逆になっている.

5 モデルの応用例

モデルの応用例として, モデルを用いたタスクの識別とシステムへの応用例として次発話予測を試みる.

5.1 タスクの識別

クロスワードパズルタスク対話, スケジューリングタスク対話それぞれから学習用として, 任意の 4 対話を選びそれぞれの Ergodic HMM を求める. この学習をそれぞれのタスクで 2 set ずつ行なった. また学習に用いなかったそれぞれ 4 対話, 2 対話を用いて, 求めた Ergodic HMM によるタスクの識別実験を行なった. 表 8 に識別結果を示す. 表中の "話者無", "話者有" は話者交替情報を利用したか否かを表し, "話題毎" とは 1 つの話題についての対話を 1 つの発話ラベル遷移列を使用したことを表している.

表 8: タスク識別結果

モデル	クロスワードパズルタスク				スケジューリングタスク			
	対話全体		話題毎		対話全体		話題毎	
	話者無	話者有	話者無	話者有	話者無	話者有	話者無	話者有
状態 2	93.8%	100%	72.7%	59.0%	87.5%	0.0%	15.7%	91.2%
状態 3	100 %	100%	87.6%	68.8%	75.0%	37.5%	8.8%	88.2%
状態 4	100 %	100%	78.4%	53.3%	87.5%	25.0%	20.6%	97.1%
状態 5	87.5%	100%	84.3%	81.0%	75.0%	0.0%	14.7%	64.7%
状態 6	100 %	100%	54.8%	64.4%	37.5%	37.5%	55.9%	88.2%
状態 7	100 %	100%	84.5%	80.3%	75.0%	0.0%	23.5%	50.0%

表8よりクロスワードパズルタスク対話の識別はかなり可能であることがわかる．この識別実験ではスケジューリングタスク対話の識別に用いたデータが少なかったことが識別率に影響していると考えられる．

5.2 次発話予測への応用

対話モデルのシステムへの利用を考える時，次発話予測への応用が考えられる．ここでは，第5.1節で求めた Ergodic HMM を利用し，オープンデータを用いて次発話ラベルの予測を行なった．なお話者交替情報を利用したモデルでは話者交替の有無と発話ラベルがともに予測できた場合を正解としている．

クロスワードパズルタスク対話での次発話ラベル予測結果を表9～12に，スケジューリングタスク対話での次発話ラベル予測結果を表13～16に示す．

表9～16よりどの状態数のモデルが次発話予測に最適かは一意にはいえないことがわかる．タスクや条件に合わせて最適な状態数を選ぶことにより次発話ラベル予測結果を音声対話システムの音声認識の際の重み付け等に利用することが可能であると考えられる．

6 まとめ

音声の利点である人間らしさやフレンドリーで協調的な対話をシステムに実現させるためにはタスクに依存しない発話も考慮して，対話全体をモデル化する必要があると考えられる．このような観点から2つのタスクによる対話データに対して Ergodic HMM を用いてそれぞれ

①：1つの対話全体を1つの発話ラベル遷移と考慮

②：対話内の話題の転換点を考慮して1つの話題内の対話を1つの発話ラベル遷移と考慮

③：①に発話権の観点から話者交替情報を付加

④：②に発話権の観点から話者交替情報を付加

の4通りの対話の状態変化のモデル化を行なった．

また，応用例としてモデルを用いたタスクの識別とシステムへの応用例として次発話予測を試みた．今後は対象とするタスクを増やすとともにそれぞれのタスク内でのデータを増やしさらにモデルの応用を検討していく必要がある．

参考文献

[1] 森川 恵美, 中里 収, 田中 修一, 白井 克彦: “協調問題解決におけるコーパスに基づく対話モデル”, 人工知能学会研究会資料, SIG-J-9501-3, pp.63-69, 1995.

[2] 中里 収, 亀山 晋, 田中 修一, 白井 克彦: “協調作業における対話データの収録と分析”, 音講論 1-Q-8, pp.157-158, 10-11 月, 1994.

[3] 平成 6,7 年度, 文部省重点領域研究「音声対話」音声対話コーパス WG, 対話音声コーパス Vol. 1,3”.

[4] 樽松 明, 中筋 知己: “スケジューリングタスクにおける自由発話音声の特徴”, 1996 年電子情報通信学会総合大会, SD4-2, pp.329-330, 3 月, 1996.

[5] 森川 恵美, 中里 収, 田中 修一, 白井 克彦: “協調問題解決における対話モデル”, 人工知能学会全国大会, pp.525-528, 1995

[6] 北 研二, 福井 義和, 永田 昌明, 森元 逞: “発話タイプ付きコーパスを用いた統計的対話モデルの自動生成”, 人工知能学会研究会資料, SIG-SLUD-9503-8(02/09), pp.47-54.

[7] Masaaki Nagata, Tsuyoshi Morimoto: “An Information-Theoretic Model of Discourse for Next Utterance Type Prediction”, Trans. of IPSJ, Vol.35 No.6, pp.1050-1061, 1994.

表 9: 次発話ラベル予測結果%(クロスワードパズルタスク:対話全体)

k-best	状態数					
	2	3	4	5	6	7
k=1	30.0	26.4	29.4	25.5	29.4	33.3
k=2	41.2	43.2	43.8	40.4	37.2	43.4
k=3	60.1	52.9	50.3	49.3	48.7	53.7
k=4	67.3	60.6	58.8	59.5	61.8	64.9
k=5	76.7	72.1	68.3	70.8	71.8	74.8

表 10: 次発話ラベル予測結果%(クロスワードパズルタスク:対話全体, 話者)

k-best	状態数					
	2	3	4	5	6	7
k=1	19.4	16.8	14.2	12.2	11.0	5.0
k=2	24.9	23.2	20.3	21.8	21.8	12.4
k=3	34.0	34.8	26.0	32.8	30.0	16.6
k=4	39.2	41.6	29.7	40.6	33.9	18.9
k=5	41.6	45.9	35.8	45.4	39.0	25.2

表 11: 次発話ラベル予測結果%(クロスワードパズルタスク:話題毎)

k-best	状態数					
	2	3	4	5	6	7
k=1	14.9	19.6	31.1	12.8	26.7	35.0
k=2	56.4	40.8	43.8	33.0	38.8	46.1
k=3	64.0	60.4	55.9	42.1	59.9	55.9
k=4	65.4	66.9	64.0	49.1	75.8	65.7
k=5	74.3	72.0	75.9	65.0	83.3	72.1

表 12: 次発話ラベル予測結果%(クロスワードパズルタスク:話題毎, 話者)

k-best	状態数					
	2	3	4	5	6	7
k=1	12.3	17.8	10.2	9.3	7.5	8.2
k=2	25.5	25.2	24.4	21.3	19.2	22.4
k=3	29.0	30.2	31.4	33.5	27.7	29.4
k=4	34.8	34.5	35.0	39.7	32.8	34.5
k=5	38.4	40.6	39.4	42.7	38.5	39.5

表 13: 次発話ラベル予測結果%(スケジューリングタスク:対話全体)

k-best	状態数					
	2	3	4	5	6	7
k=1	26.5	21.0	19.6	15.1	19.2	16.0
k=2	45.2	39.3	34.7	32.9	46.1	37.0
k=3	55.7	49.8	48.4	47.5	59.8	49.8
k=4	69.4	62.1	58.4	60.7	72.1	63.9
k=5	85.4	68.0	72.6	70.7	76.7	78.5

表 14: 次発話ラベル予測結果%(スケジューリングタスク:対話全体, 話者)

k-best	状態数					
	2	3	4	5	6	7
k=1	18.7	21.9	16.4	1.8	21.5	7.8
k=2	29.7	36.1	31.1	7.3	32.0	13.2
k=3	40.2	41.6	37.4	13.7	41.1	31.1
k=4	45.7	48.4	42.0	27.9	52.5	37.4
k=5	51.6	51.6	50.7	30.1	55.8	43.4

表 15: 次発話ラベル予測結果%(スケジューリングタスク:話題毎)

k-best	状態数					
	2	3	4	5	6	7
k=1	30.5	20.4	22.2	19.4	25.7	25.7
k=2	42.8	37.9	36.9	40.3	38.8	41.7
k=3	73.1	53.4	56.8	61.2	53.9	51.5
k=4	78.3	57.3	65.0	75.2	63.6	63.1
k=5	87.2	72.3	77.2	84.5	72.3	79.6

表 16: 次発話ラベル予測結果%(スケジューリングタスク:話題毎, 話者)

k-best	状態数					
	2	3	4	5	6	7
k=1	13.1	6.8	21.4	16.0	14.1	17.0
k=2	25.7	17.5	29.1	30.1	25.2	26.2
k=3	34.0	34.0	37.9	37.9	32.5	38.3
k=4	43.2	58.7	47.1	48.1	44.2	44.2
k=5	49.0	61.7	54.9	57.8	48.1	55.8