音声言語情報処理 5－9
（1995. 2. 4）

音声対話システムの構成法とユーザ発話の関係の模擬実験による評価

山本 誠治　中川 聖一

豊橋技術科学大学　情報工学系

概要

近年様々なタスクを扱った多くの音声対話システムが試作されている。しかしそれらのシステムの評価方法はシステム構成やタスクに依存したものが多く、異なる機関で設計されたシステムに対して公平な評価が行なえるような明確な方法はまだ確立されていない。

本報告では、“Wizard of Oz” 法でインプリメントしたシステムを用いて、タスク、対話方法、応答文（規則合成音声か録音編集音声か）、被験者の違いによる被験者の振舞い、主観の違いを検討する。また音声対話システムの一般的な評価方法についての一考察を提案し、本実験での結果をもとにその妥当性を検討する。

An evaluation of the relationship between the structure of spoken dialog systems and user utterances through simulated experiments

Seiji Yamamoto and Seiichi Nakagawa

Department of Information and Computer Sciences
Toyohashi University of Technology

Abstract

Recently various spoken dialog systems have been developed. But traditional methods of evaluating such systems depend on the systems' structures and tasks they handle; and as of now, there has been no decisive way of fairly evaluating systems developed in different laboratories.

In this paper, through simulated experiments using spoken dialog systems implemented by "Wizard of Oz" method, we examine subjects' behavior and subjectivity when tasks, system structures, system responses and subject types are changed. Furthermore, we propose an evaluation method for spoken dialog systems and examine its validity using the results from the experiments.

1 はじめに

近年様々なタスクを扱った多くの音声対話システムが試作されている。しかしそれらのシステムの評価方法はシステム構成やタスクに依存したものが多く、異なる機関で設計されたシステムに対して公平な評価が行なえるような明確な方法はまだ確立されていない。

一般に人間—人間の対話に比べて人間—機械との対話では、人間の発話は形式的、幼児的、文字言語的、機械的になると言われている [1]。R.W.Smith[2] は、システム及びユーザが主導権を持つ 2 つのシステムの比較を行い、ユーザが主導権を持つシステムの方が 1 対話当たりの発話数が少なく、発話が長くなるという結果を示している。

本報告では、"Wizard of Oz" 法 [1] でインプリメントしたシステムを用いて、タスク、対話方法、応答文(規則合成音声か録音編集音声か)、被験者の違いによる被験者の振舞い、主観の違いを検討する。また音声対話システムの一般的な評価方法についての一考察を提案し、本実験での結果をもとにその妥当性を検討する。

本稿では、2 節で現在我々が考えている音声対話システムの評価方法とその問題点について報告する。また 3 節で"Wizard of Oz 法" を用いて行った評価実験の概要とその結果について報告する。そして 4 節で本研究のまとめと今後の課題について報告する。

2 システムの評価法と問題点

ある音声対話システムの性能を考える場合、システムのカバーレイジ、処理時間、正解率 (タスク達成率) などは、システムの性能を知る 1 つの方法である。もし全く同じタスクを扱うシステム同士の比較を行なうのであればこのようなメジャーは有効であるが、異なるタスクを扱うシステムを比較する場合、システムのカバーレイジ、処理時間、正解率を単に比較することは出来ない。我々は異なるタスク間の比較のために、タスクの複雑さというメジャーを考えている。

まず最初に考えた方法は、タスクの複雑さを意味表現の数で表すという方法である。これは、対話の仕方や言い回しなどに影響されないという利点がある。しかしどのような意味表現を用いるのか、またどのようにして意味表現の数を求めるのか、というような様々な問題点があげられる。そこで別の方法として、タスク内で現れる対話数でその複雑さを表そうと検討した。タスクの複雑さを対話数で考えるのであれば、タスクは対話の集合として簡単に定義できる。しかし表層レベルの対話で考えると、言い回しがタスクの複雑さに考慮され (我々は、言い回しはシステム側の問題と見ている)、深層レベルの対話で考えると先ほどの意味表現の場合と同じような問題点が考えられる。その後種々の検討を重ねた結果、タスクの複雑さを次のように考えることにした。

タスクの複雑さとは、"システムがそのタスクを処理するのがどれほど困難か" という度合を表すメジャーである。曖昧さが大きい対話ほどシステムにとっては処理が困難であると考えれば、タスク内の対話で得られる情報量が大きいほどタスクは複雑であるといえる。そのため我々はタスクを部分対話 (1 つの情報が得られる最小単位の対話) の集合で定義し、タスクの複雑さをそのタスクで得られる情報量で定義した。

<タスクの定義>

- タスクとは "ある情報を得るとそれに対する情報を返す" というような部分対話の集合である。
- タスクはシステムと独立である。

<タスクの複雑さの定義>

- ある部分対話の出力 (応答) Y_iが得られる直前の入力 X_iの不確かさの程度を $H(X_i)$、出力 Y_i が得られた直後の入力 X_iの不確かさの程度を $H(X_i|Y_i)$ とするとタスクの複雑さ TC は、

$$TC = \sum_{i=1}^{I} \{H(X_i) - H(X_i|Y_i)\} \qquad (1)$$

  で求められる。
- タスクの複雑さはシステムと独立である。

このようにタスクの複雑さを相互情報量で表すことにより、システムの性能を単位時間、単位ターン数当たりに得られる情報量として以下のように定義した (TC を部分対話の数で正規化する定義もありうる [3])。

<システムの評価法>

- 目的とするシステムで扱っているタスクの複雑さを TC、正解率 (達成率) を CT、全ての対話が終了するまでの時間を t、ターン数を T_n とすると、システムの性能 SP は、

$$SP = \frac{TC \cdot CT}{t \cdot T_n} \qquad (2)$$

  で表される。

(2) 式は、複雑なタスクを扱っているほど、正解率が良いほど、所要時間が短いほど、そしてターン数が短いほど優れたシステムであるということを意味する。

しかし、この定義の方法もいくつかの問題点を抱えている。それは、例えば、部分対話をどのようにして定義するのかということや、その情報量をどのようにして求めるのかということなどである。例えばホテルの名前が分かった時に得られる相互情報量は、事前の可能なホテル数に依存してしまう。更にその数もユーザによって異なる場合もある。この定義の実験的評価は 3.2.5 節で述べる。

3 Wizard of Oz 法を用いた対話システムの評価実験

この節では、"Wizard of Oz 法" を用いた評価実験について述べる。この実験の目的は、音声対話システムにおける対話の傾向を知ることである。また、先に示したシステムの評価方法の妥当性を検討することも、実験の目的の1つである。

本実験では、タスク、対話方法、応答文（規則合成音声か録音編集音声か）、被験者のグループなどを変化させて実験を行った。各実験の実験方法と結果を以下に示す。

3.1 実験方法

3.1.1 実験1

実験1では、システムの対話方法の違いや、タスクの違いによる被験者の振舞い、主観などを検討するため、タスクおよびシステムの構成法を次のように設定し、これらの組合せで実験を行なった。また本大学の学生を被験者とし、合計18人（6人ずつ3組）で実験を行った。その結果、1329発話の対話データを収集した（文献[3]を参照）。

1. タスクは、"富士五湖周辺の宿泊施設案内" と "航空座席予約" の2つを考える。
2. システムの構成法としては、システム主導型とユーザ主導型を考える。

なお実験1で用いる音声対話システムでは、入力音声の認識から応答文の作成までを人間が行ない、応答の出力だけを機械（規則による音声合成システム）が行った。また音声認識率を100%と仮定し、システムの応答できる許容範囲外の質問には、" それはシステムの範囲外のことなので分かりません" と応答した。

システムの使用方法や実験に対する指示を正確にかつ各被験者に対して同じように説明するため、実験1では、先ず文書により被験者に必要な指示を与えた（2度読む）。この指示には、システムにどういう条件を指定できるのか（宿泊施設案内タスクでは、宿泊施設の種類、湖、部屋の種類、夕食の有無、その他を、航空座席予約タスクでは、出発地点と目的地、ひにち、航空会社、出発時間、その他を具体的に列挙している）と、どういう課題を達成して欲しいのか、そしてシステムを使用する上での注意事項を示した。その後でビデオによりシステムの使用方法（実際の実行例の一部）を示し、以下に示すシナリオに沿ってシステムを使用してもらった。アンケートは実験の後に答えてもらい、システムにより得られた情報は、別紙に記入してもらった。なおシステム主導型とユーザ主導型のシステムでは異なる課題で実験を行ったが、どちらのシステムでも同じ課題を実行できる（つまり2つのシステムでは同じ情報量が得られる）。

＜シナリオ＞

- 宿泊施設案内では2泊3日の旅行を行なうと仮定し、その時に利用したい宿泊施設について調べる。
- 航空座席予約では、2ヶ所の都市訪問のために航空チケットを予約をする。

被験者に示した課題の詳細は以下に示す通りである。

＜被験者に与えた課題＞

1. 宿泊施設案内タスクでシステム主導型の場合
   - 1日目と2日目の宿泊先は違う方が良い。
   - 1日目の夕食はある方が良い。
   - 最初の宿泊施設への行き方は聞いておく方が良い。
   - 2日の内、どちらかはペンションが良い。
2. 宿泊施設案内タスクでユーザ主導型の場合
   - 1日目は7000円くらいのペンションが良い。
   - 1日目は河口湖周辺が良い。
   - 2日目はプール付きの宿泊施設が良い。
   - 2日目は山中湖のホテルが良い。
3. 航空座席予約タスクでシステム主導型の場合
   - 最初の都市はニューヨークかワシントンで、2番目の都市はロサンゼルスかサンフランシスコのどちらかである。
   - 日本から最初の都市への飛行機は、ノンストップ便を希望する。
   - 2番目の都市から日本への飛行機は、一番早い便を希望する。
4. 航空座席予約タスクでユーザ主導型の場合
   - 最初の都市はロサンゼルスかサンフランシスコで、2番目の都市はニューヨークかワシントンのどちらかである。
   - 日本から最初の都市への飛行機は、18時頃の便を希望する。
   - 2番目の都市から日本への飛行機は、食事つきの便を希望する。

同じタスクを扱うシステムで対話方法を変えて実験を行なう場合、後から使用する対話方法の方が有利になる可能性がある。そこでシステムを使用する順序（ユーザ主導型からするか、システム主導型からするか）を変えて実験を行なった。

3.1.2 実験2

実験2では、被験者として工学系でない一般女性（本大学とは無関係の女性4人）を選んで実験を行った。タスクは富士五湖周辺の宿泊施設案内で、応答文として規則合成音声を使用した（被験者以外は実験1と同じ）。その結果、275発話の対話データを収集した。

3.1.3 実験3

実験3ではタスク及びシステムの構成法を以下のように設定し、これらの組合せで実験を行った。

1. タスクは、"富士五湖周辺の宿泊施設案内" と "豊橋市内の飲食店の登録" の2つを考える。
2. システムの構成法としては、システム主導型とユーザ主導型を考える。

飲食店の登録システムを使用する被験者への指示には、どういう情報をシステムに登録できるのか（飲食店の名前、場所、店の種類、メニューと料金、その他を具体的に列挙している）と、どのような課題を達成して欲しいのか、そしてシステムを使用する上での注意事項を示した。飲食店の登録システムで被験者に与えたシナリオと課題を以下に示す。

＜シナリオ＞

- 豊橋市内の飲食店に関する情報を2軒分登録してもらう。

＜被験者に与えた課題＞

1. 飲食店の登録タスクでシステム主導型の場合
   - 登録する飲食店は、豊橋市内の飲食店に限ります。
   - 飲食店を2軒登録して下さい。
   - 飲食店の名前は必ず登録して下さい。
2. 飲食店の登録タスクでユーザ主導型の場合
   - システム主導型のシステムで登録した飲食店とは異なる飲食店を2軒登録して下さい。
   - それ以外はシステム主導型の場合と同じです。

被験者は本大学の学生で、合計11人（タスクを固定し対話方法を変化させる実験は4人ずつ2組、2つのタスクで実験を行う場合は3人）で実験を行った。なお実験3では応答文として録音編集音声を使用した。その結果、686発話の対話データを収集した。

なお、飲食店の登録システムを使用する被験者には、システムにより得られた情報を記入する別紙を渡さなかった（システムから情報を引き出すタスクではないため）。それ以外は実験1と同じである。

3.2 実験結果

ユーザ主導型システム及びシステム主導型システムにおいて発声されうる語彙数に関する検討は、文献[4]を参照されたい。

3.2.1 ユーザの発話数と user_time についての結果

実験1〜3で得られた対話データにおけるユーザの平均発話数と user_time ave. を表1、2に示す[3]。なお、user_time とは、システムの応答が終了してから被験者が質問を終了するまでの時間である。つまり、user_time には被験者が質問を考えている時間も含まれる。なお、実験1、2、3での課題達成率はそれぞれ83.3%、81.3%、97.5%であった（3.2.3節詳述）。

表から以下のことが言える。

1. システム主導型システムの方がユーザ主導型システムよりユーザの平均発話数が多くなる。
2. ユーザ主導型システムの方がシステム主導型システムより user_time ave. が長くなる。
3. 一般のユーザは、システムの許容範囲外の質問文を多くする。

結果1は、Ronnie W. Smith[2] の結果と一致する。1のような結果になったのは、ユーザ主導型のシステムでは被験者が複数の項目を同時に入力していたためである（基本的にシステム主導型のシステムでは、項目を1つずつ入力していく）。結果2は、被験者が発話を開始するまでの時間が長くなりがちであったためである。つまり被験者にとっては、ユーザ主導型のシステムの方が何をどのように質問すれば良いかを考えるために質問の作成に時間がかかると言える。

実験2での被験者（一般の女性）は実験1、3の被験者に比べシステムの受理・理解可能な入力文の範囲外の質問を多く行った。具体的には実験1（宿泊施設案内タスクで対話方法を変化させた実験の場合）ではシステムの許容範囲外の質問が1人当たり平均0.5文未満であるのに対し、実験2では1人当たり平均4文であった。

1人だけであるが本大学の学生（女性）で同じ実験を行った所、システムの許容範囲外の質問は1文だけであった（つまり上述の原因は男女の差ではない）。被験者の数が少ないのではっきりとは言えないが、工学系でない一般の人が音声対話システムを使用する場合、システムの許容範囲外の質問を行う割合が多いと言える。

また、実験2では対話方法の違いによる平均発話数の差が小さい。これはシステム主導型の対話では聞かなかったことをユーザ主導型の対話で聞いていることなどが原因である。このような結果から、計算機のインターフェースに対してある程度知識のある人とそうでない人との間には、その振舞いなどに違いがあると言える。

表 1: ユーザの平均発話数

実験	実験1				実験2		実験3			
タスク	宿泊施設案内		航空座席予約		宿泊施設案内		宿泊施設案内		飲食店の登録	
sys 主導/usr 主導	sys	usr	sys	usr	sys	usr	sys	usr	sys	usr
平均発話数(文)	33.3	16.3	53.0	31.7	38.9	30.0	40.5	24.5	37.8	16.8
user_time ave.(秒)	8.5	15.0	5.7	11.8	9.0	23.3	6.5	15.3	4.5	7.5

表 2: システムの使用順序を変化させた場合の user_time ave.(秒)

実験		実験1				実験2		実験3			
タスク		宿泊施設案内		航空座席予約		宿泊施設案内		宿泊施設案内		飲食店の登録	
実験の順序		s→u	u→s	s→u	u→s	s→u	u→s	s→u	u→s	s→u	u→s
time	sys 主導	5.7	11.3	6.7	4.7	7.5	10.5	5.5	7.5	4.0	5.0
	usr 主導	8.7	21.3	11.7	12	18.0	28.5	10.5	20.0	7.0	8.0
time の差		3.0	10.0	5.0	7.3	10.5	18.0	5.0	12.5	3.0	3.0

なお実験 1、3 の両方で異なるタスクを扱うシステムに対しても同様に検討してみたが(対話方法はシステム主導型)、ユーザの平均発話数、user_time ave. ともほぼ同じような結果になり、タスクの違いによる典型的な違いは見られなかった。このようにユーザの平均発話数に変化が見られなかった原因は、宿泊施設案内タスクと航空座席予約タスクでシステムに指定する項目の数と、飲食店の登録タスクでシステムに登録する項目の数にそれほど差がなかったためである。また user_time ave. については、今回用いたタスクでは、タスクの違いよりも対話方法に強く影響されたと言える。

表2 中の time は、user_time ave. を示している。また s→u は、システム主導型の後にユーザ主導型のシステムを使用したことを、u→s は、その逆の場合を示している。表2 から、システム主導型の後にユーザ主導型のシステムを使用した場合の user_time ave. の差は、その逆の順序で使用した場合よりも小さくなる。つまり、システムに対する被験者の慣れが見られる(飲食店の登録システムではこのような結果が得られなかった)。

3.2.2 実験直後のアンケートからの結果

本実験では、被験者が2つのシステムを使用した後で、1つ目のアンケートをとった。このアンケートでは、合計15項目についての質問を行っている。実験1 でのアンケートの結果は、以下に示す通りである。なお本実験では1人に2つのシステムを使用してもらったが、両方のシステムとも全てが機械である(人間が介在していない)と感じた被験者は、33 人中 19 人であった。それ以外の人は、どちらかのシステムもしくは両方のシステムに人間が一部介在しているのではないかと答えていた。

1. すぐ使えるという点では、システムに対して何も知らない被験者でも、受身的に対話を進められるのでシステム主導型の方がよい。
2. ユーザ主導型の方は被験者の方から自由に質問できるので、ある程度慣れればユーザ主導型の方が使い易い(被験者の予想)。
3. ユーザ主導型の方がシステム主導型よりも知的である(人間に近い)。
4. ユーザ主導型の方が、自分の聞きたいことが聞きたい時に聞ける。
5. 被験者がどういうつもりでシステムに話しかけたのかということと、対話中に現われる間投詞数、言い直し数とは相関がある。つまり話し方を機械が理解し易いように意識して話した場合は、間投詞、言い直しの数は少ないが、そうでない場合は前者よりも多くなる。

なお、これはタスクを固定して異なる対話方法のシステムで実験を行った場合の結果である。対話方法をシステム主導型に固定して異なるタスクを扱うシステムで実験を行った場合のアンケートからは、5 以外には顕著な違いは見られなかった。

実験2 のアンケートでは、上記 4、5 と以下に示す 6〜8 の結果が得られた。

6. システム主導型の方がユーザ主導型よりも快適である。
7. システム主導型の方がユーザ主導型よりも知的である(人間に近い)。
8. システム主導型の方がユーザ主導型よりも優れている。

6〜8 の結果は、実験1、3 と異なる(実験1、3 では、6、8 についてはどちらかというとユーザ主導型の方を被験者は好んでいた)。原因としては、実験2 の被験者がユーザ主導型のシステムでシステムの許容範囲外の質問を多く行ったことがあげられる(この場合、被験者が望む答えが得られないので、システム主導型の方を優れていると感じるのは自然である)。つまり被験者が計算機への入力に慣れていないのが原因と思われ、一般のユーザを対象とした対話システムの構築には、対話ストラテジーが重要であることを示している。

実験3 のアンケートでは、タスクを固定して対話方法を変えた場合は、上記 1、2、3、5 と以下に示す 9

の結果が得られた（ただし宿泊施設案内タスクで実験を行った4人の被験者の内、2人は4と同じように答えた)。

9. ユーザ主導型では、自分が登録したいことを登録したい時に登録出来る（飲食店の登録タスク)。

対話方法を固定して、異なるタスクを扱うシステムで実験を行った場合の結果は、上記5と以下に示す10〜13である。

10. 飲食店の登録システムの方が使い易い。
11. 宿泊施設案内システムの方が知的である。
12. 宿泊施設案内システムの方が複雑な対話を扱っている。
13. 宿泊施設案内システムの方が優れている。

宿泊施設案内システムでは、システムから様々な情報が得られるが、飲食店の登録システムは、被験者から情報をシステムに登録するだけのシステムなので、10〜13のような結果になった（アンケートより)。

3.2.3 課題達成率と別紙への情報の書き込み状況についての結果

本実験では、被験者にいくつかの課題を与え、システムを使用して得られた情報を別紙に記入してもらった。実験1、2、3での課題達成率はそれぞれ83.3%、81.3%、97.5%で、被験者が一般の人の場合が一番悪かった。実験1、2での課題達成率が実験3に比べ低かったのは、システム側の応答ミスというよりは被験者側のミスである（1つ目と2つ目のシステムの課題の混同など。例えば、山中湖と河口湖の勘違いなどがあった)。

なお飲食店の登録タスクでの課題達成率は100%であった（他のタスクに比べ課題が簡単であったため)。

3.2.4 聞き直しと間投詞、言い直しについての結果

本実験では、システムの応答を聞き直したい時には、"もう一度" とシステムに入力するように被験者に指示を与えた。聞き直しについては以下のことが言える。

1. 宿泊施設案内タスクの方が航空座席予約タスクよりも聞き直しが多い（1発話当り7.7%対2.6%)。
2. 一般の人が被験者の場合、本大学の学生の場合より聞き直しが多い（1発話当り9.8%対7.7%)。
3. 飲食店の登録タスクでは聞き直しはほとんどない（今回の実験では0であった)。
4. 応答文として規則合成音声を用いると、録音編集音声を用いる場合よりも聞き直しが多い（1発話当り7.7%対4.1%)。

1のような結果になった理由は、宿泊施設案内タスクの方が航空座席予約タスクよりも聞きとりにくい文（前者では宿泊施設名を列挙した文や行き方を示した文、後者では便名や時間などを列挙している文）が多かったからである。結果2については、ある程度予想された結果である。何故なら、一般の人よりも本大学の学生の方が何らかのシステムに触れる機会が多いため、システムに対する知識があり、かつシステムの使用に対する緊張、不安などが少ないと考えられるためである。また結果3については、タスクの性質が大きく影響している（飲食店の登録タスクではシステムから情報を聞くことはほとんどないので、聞き直しをする必要がなかった)。結果4もある程度予想された結果である。実際、応答文として規則合成音声を使用した場合、録音編集音声では考えられないような短い応答に対して聞き直しが行われている場合がいくつか見られた。

実験1〜3により得られた書き起こし文中に現われる間投詞、言い直しの割合についての結果を表3に示す。

表3中の間投詞と言い直しは1発話当りの数を示しているため、これらを比較する場合は1発話当りの文節数を考慮する必要がある。宿泊施設案内タスクで対話方法を変化させた実験（表3の左）では、システム主導型とユーザ主導型の1発話当りの文節数比が1.51(=2.71/1.79)で、1発話当たりの間投詞数の比が2.77（=19.4/7.0）であるため、宿泊施設案内タスクでは、ユーザ主導型の対話の方が間投詞は多いと言える。以下同様に考えると、表3(実験1)から次のことが言える。

1. システム主導型の対話よりもユーザ主導型の対話の方が間投詞が多い。
2. システム主導型の対話よりもユーザ主導型の対話の方が言い直しが多い。
3. 宿泊施設案内タスクの方が航空座席予約タスクよりも間投詞が多い。
4. 宿泊施設案内タスクの方が航空座席予約タスクよりも言い直しが多い。

結果1は、ユーザ主導型の対話の方がシステム主導型の対話よりも被験者が質問を考える時間が長かったためである（表1参照)。結果2は、ユーザ主導型の対話の方が比較的長い文になりがちであったためである（表3の文節の欄及び文献[4]を参照)。結果3、4も結果1、2と同様で、宿泊施設案内タスクの方が user_time ave. が長く、1発話当たりの文節数が多かったためである。

表3(実験2)では、言い直しについては結果2と同じ傾向になっているが、間投詞については、1発話当たりの文節数を考慮すると結果1とは逆の結果になっている。また表3(実験3)では、飲食店の登録タスクにおける言い直し以外は1、2と同じ結果になっている。

表 3: 間投詞と言い直しの割合（%）

実験	実験 1						実験 2		実験 3					
タスク	宿泊施設		航空座席		航空	宿泊	宿泊施設		宿泊施設		登録		宿泊	登録
sys 主導 /usr 主導	sys	user	sys	user	sys	sys	sys	user	sys	user	sys	user	sys	sys
発話数	200	98	318	190	262	261	155	120	162	98	151	67	102	106
文節数/発話	1.79	2.71	1.38	2.92	1.46	1.71	1.58	2.88	1.93	3.17	1.58	3.60	1.56	1.54
間投詞	7.0	19.4	3.8	15.3	3.8	6.1	10.3	15.8	6.8	33.7	6.6	46.3	1.0	0.9
言い直し	2.0	3.1	1.3	7.4	1.1	1.9	1.9	5.8	1.9	9.2	2.6	0.0	2.9	0.0

実験 2、実験 3 の一部の結果が実験 1 の結果と異なっているが、user_time ave.、1 発話当たりの文節数を考慮するとこれは個人差によるものと言える。

表 3 の結果と小林ら [5]、上條ら [6] の分析結果とを比較すると、小林ら、上條ら、本実験の順で間投詞、言い直しが多い。つまり、本実験は他の 2 つの実験よりも機械を相手にした対話に近いと言える。

本実験の結果が小林らの結果に比べ間投詞や言い直しが少ないのは、小林らの結果が人間同士の自然な対話であるのに対し、本実験は人間対機械の対話を扱っているためである。また上條らの結果よりも間投詞や言い直しが少ないのは、タスクの違いや被験者のシステムへの接し方の違いである（アンケートでは、33 人の被験者の内 32 人が話し方を機械が理解し易いように意識したと答えている。ただし、話し方を機械が理解し易いように意識して、かつ間投詞や言い直しを行わないように意識した被験者は、32 人中 14 人であった）。

3.2.5 システムの性能（SP）についての検討

本実験では、ユーザが得ようとした情報に対するシステムの応答の正解率（達成率）を 100%として実験を行ったため、(2) 式より SP はタスクの複雑さ TC、所要時間 t、ターン数 T_n によって求められる（実際は正解率（達成率）は 100%ではなかったが、99%以上はあった）。所要時間 t とターン数 T_n は簡単に求められるが、2 節でも述べた通り部分対話とその情報量を求めるのが困難である。そのため、今回は各発話によって得られる情報量をアンケートによって被験者自身に決定してもらった。このアンケートには、各被験者が実際に行った対話と被験者全員に共通のスタンダードな対話を示し、各被験者に 1 から 5 の 5 段階で数字をつけてもらった。なお情報量という言葉を用いると数字をつけるのが難しいという被験者の意見があったため、システムの応答に対する被験者の応答のバリエーションの数が多いほど大きい数字をつけてもらった（被験者の応答のバリエーションの数が多いほど、システムにとって難しいタスクであると考えられる）。しかし各被験者が割り当てる数字には個人差があるため、アンケート中のスタンダードな対話につけられた数字を元に正規化を行って TC を求め、(2) 式を用いて SP を計算した。その計算結果と実験直後に行ったアンケートの結果（" どちらのシステムの方が優れているか" という項目）を比較した。なお計算結果によりどちらのシステムの方が優れているのかを比較する場合は、計算結果の数値に 1 割以上の差がある方を優れていると判断し、差が 1 割未満の場合は同程度と判断した。

実験 1 のアンケートでは、18 人中 13 人がどちらかのシステムの方が優れていると答えた（残りの 5 人は同じくらいと答えた）が、その内 10 人（76.9%）の被験者の主観と計算結果が一致した。

実験 2 では 4 人の被験者の内 1 人（25.0%）の被験者の主観と計算結果が一致したに過ぎなかった。これはランダムに選んだ場合（50%）より悪い。このような結果になった理由としては、実験 2 での被験者が実験 1、3 に比べてシステムの許容範囲外の質問を多く行ったことがあげられる。そのためユーザ主導型の対話を扱うシステムに対する評価が低くなった。従って実験 2 での結果から 2 節で示したシステムの評価方法の妥当性を議論することは出来ない。

実験 3 においても同じように比較した。タスクを固定して対話方法を変えて実験を行った場合は、8 人の被験者の内 6 人（75.0%）の被験者の主観と計算結果が一致した。しかし、対話方法をシステム主導型に固定して異なるタスクで実験を行った場合は、3 人の被験者の主観と計算結果とは全く逆になった（被験者は宿泊施設案内タスクの方が優れていると答えた）。これは、飲食店の登録タスクではシステムに情報を登録することだけしか出来ないということが大きく影響している（アンケートより）。

ここで示した計算結果は理論的に求めた結果ではない（被験者に割り当ててもらった数字をもとにタスクの複雑さを求めている）が、実験 1 と実験 3 の一部の被験者に対して 75.0%以上の一致率が得られているため有効であると言える。

4 まとめと今後の課題

本報告では、"Wizard of Oz 法" を用いてインプリメントしたシステムを用いて、タスク、対話方法、被験者、応答文の違いによる被験者の振舞や主観と、書き起こしデータ中に現われる様々な現象について考察した。その結果は、以下に示す通りである。

1. ユーザ主導型のシステムの利点

   基本的に被験者は、ユーザ主導型のシステムの方が優れていると感じる。何故ならシステムへの質問や登録の順序を被験者自身が選択でき、かつ複数の項目を同時に指定できるためである。また、ある課題を達成するまでの平均発話数が少なくなることも要因の1つである。

2. ユーザ主導型のシステムの欠点

   ユーザ主導型のシステムは、ある程度タスクに慣れるまでは少し使いにくい（被験者の予想）。何故なら質問を考えるまでに時間が比較的かかるため（何をどのように質問すれば良いかが不明なため）である。つまり初心者にとっては、システム主導の方が使い勝手が良い。また被験者に選択権があるため、システム主導型のシステムに比べるとシステムの許容範囲外の応答がなされる可能性が高くなる。一般の人が被験者の場合は、この傾向が一層強くなる（システムへの聞き直し回数も多くなる）。更にシステムを設計する立場から考えると、異なり語彙数、間投詞、言い直しの割合が多くなるのが問題である [4]。

3. システム主導型のシステムの利点

   システムの方から積極的に質問を行ってくるため、システムについて何も知らない被験者でも受身的に対話を進められるためすぐに使えるようになる。また、システムを設計する場合、1発話当たりの文節数が少ないのでその分だけ音声認識や構文解析が容易になる。

4. システム主導型のシステムの欠点

   システムに主導権があるため、被験者から自由に質問したり登録することが基本的には出来ない。

5. 応答文が録音編集音声の場合の利点と欠点

   録音編集方式は、システムからの応答の質が向上するため、被験者の聞き直しがかなり減少する。また、被験者にも好感が持たれる。しかし、より大きなタスクを扱うシステムを作成する場合、使用される単語などの録音が困難である（全てをカバーするのは困難である）。更に、全ての応答文を文単位でシステムに登録しておくのであれば問題ないが、単語単位などで登録する場合、つなぎ目での音声の質が問題となる。

以上の結果から、対話方法や応答文の違いにはそれぞれ利点、欠点があるため、これらのトレードオフによってシステムを作成する必要がある。また対象とする被験者のグループの種類などもこれらを選択する要因の1つである。

また、異なるタスクを扱うシステム、あるいは対話方法が異なるシステムの評価方法を提案し、"Wizard of Oz 法" で得られたデータをもとにこの評価方法の妥当性を検討した。その結果、対話方法が異なるシステムの比較や傾向が同じタスクを扱うシステムの比較を行う実験では、75%以上の一致率が得られたため、このようなシステムを比較する場合は妥当であると言える。しかし、傾向が異なるタスクを扱うシステムの比較実験では良い結果は得られなかった。また被験者が工学系でない一般の人の場合も、システムの許容範囲外の質問が多かったなどの理由により、良い結果は得られなかった。しかしシステムの許容範囲外の質問を行わないようにすれば、また違った結果になると考えられる。宿泊施設のサブタスクでもユーザ主導型では語彙数は 1000～2000 以上になるのに対し、システム主導型では 200～300 程度で十分である [4]。当面は、システム主導で十分と考えられる。

今後の課題としては、部分対話の求め方や部分対話によって得られる情報量を理論的に求める方法の検討があげられる。また、飲食店の登録タスクのような傾向が異なるタスクにおいても妥当な結果が得られる様に、2節で示した評価方法を改良する必要がある。

参考文献

[1] N.M.Fraser and G.N.Gilbert：Simulating speech systems, Computer Speech and Language, Vol.5, pp.81-99(1991)

[2] Ronnie W. Smith：Spoken Variable Initiative Dialog:An Adaptable Natural-Language Interface, IEEE EXPERT, pp.45-50(Feb. 1994)

[3] 山本誠治、山本幹雄、中川聖一：音声による対話システムの評価法における一考察, 情報処理学会講演論文集, 6G-05(1994.9)

[4] 伊藤敏彦、大谷耕嗣、肥田野勝、山本幹雄、中川聖一：事前説明によるシステムへの入力発話の変化と語認識結果の人間による復元, 情報処理学会研究報告, 94-SLP-4-7(1994.12)

[5] 小林聡、山本幹雄、中川聖一：間投詞、言い直し等の出現に関する音響的特徴, 情報処理学会研究グループ資料, No.93-SLP-1-2(1993.7)

[6] 上條俊一、秋葉友良、伊藤克亘、田中穂積：音声対話データの分析と発話理解への応用, 人工知能学会研究会資料, SIG-SLUD-9402-6(1994.10)