音声言語情報処理 29－27
(1999. 12. 21)

音声対話システムの誤解に対するユーザ応答の分析

平沢 純一　宮崎 昇　中野 幹生　相川 清明

NTT研究所
〒243-0198 神奈川県厚木市森の里若宮 3-1
jun@idea.brl.ntt.co.jp

あらまし　音声対話システムが何らかの誤解を起こすことは避けられない。ユーザに特別な話し方や操作をさせることなくシステムが誤解を克服するには、システムがユーザとのやりとりの中から自らの誤解を検出し、復旧していく技術が必要となる。本稿ではまずシステムの誤解発話に対するユーザの反応を収集し、システムの誤解に対するユーザ応答の特徴を報告する。実験の結果、正しい理解に基づくシステム発話に対するユーザ応答は、内容語をほとんど含まずに肯定表現が多用され、発話長が短い傾向にあった。これに対して、誤解に基づくシステム発話に対するユーザ応答は、否定表現は多用されず、ユーザが正しい内容を繰り返す訂正発話が行われ、発話長が長い傾向にあった。これらの結果より、システムが対話の中から自らの誤解を検出できる可能性が示された。

キーワード：　音声対話システム　誤解　誤解の訂正　確認発話

Studies on How Users Respond to Misunderstandings of Spoken Dialogue Systems

Jun-ichi Hirasawa　Noboru Miyazaki
Mikio Nakano　Kiyoaki Aikawa

NTT Laboratories
3-1 Morinosato Wakamiya, Atsugi-city, Kanagawa 243-0198 Japan
jun@idea.brl.ntt.co.jp

Abstract　Conventional spoken dialogue systems cannot avoid misunderstandings. A spoken dialogue system should be able to correct its own misunderstandings in the course of dialogues without any special expressions or operations from the user. For this purpose, we conducted an experiment to investigate how the users respond to the system's misunderstandings. The results show that the user responds to the system's correct verifications with many affirmative expressions and fewer repeated content words in shorter utterances. On the other hand, when the user responds to the system's misunderstood verifications, the user hardly uses negative expressions and usually repeats his original content words in longer utterances. These observations would lead to a system that can detect and correct its own misunderstandings.

key words : spoken dialogue system, misunderstanding, error correction, verification utterance

1 はじめに

人間とコンピュータが音声対話によって何らかの仕事（タスク）を遂行できることを目指している。音声対話を用いるなら、特別な操作法を学ばなくても日常会話のような感覚で「気楽」にコンピュータを使える。しかし現状の技術ではコンピュータの音声認識や音声理解の誤りを避けられず、タスクが遂行されないことがある。そこで「気楽さ」という音声対話の利点を損なわずにタスク遂行を確実にするためには、ユーザに特別な話し方や操作をさせることなく、自然なやりとりが進行する中で音声対話システムがシステム自身の誤解を検出・復旧していく技術が急務となる。

我々が目指しているのは、音声対話システムがユーザとの対話のやりとりの中から、システム自身の理解を「正しい」ものとして確定（ground）するのか、「誤解している」として誤りを復旧するのか、を判断する技術の開発である。つまり、音声対話システムはユーザからの音声入力を処理する中で、対話が「正しい理解で進んでいる事態」と「自らの誤解による異常事態」のどちらにあるかを判断できなければならない。本稿では、システムが自らの誤りを検出・復旧する技術のために、まずシステムの誤解発話に対するユーザの反応を収集し、システムの誤解に対するユーザ応答の特徴を報告する。

音声対話システムが自らの誤解を検出する技術は実用的なシステムで重要になる要素技術であり、確認発話をすれば誤解しているかどうか判明する、という問題ではない。例えば、システムが「予約は月曜日でよろしいでしょうか？」のようなyes/no質問で確認した時に、ユーザが必ず「はい / いいえ」で応答してくれる保証はない[3]。しかしシステムからの確認の度に「予約は月曜日ですか？ハイかイイエでお答え下さい」と質問するのでは「気楽さ」という音声対話の利点を著しく損ねている。そもそも、確認発話に対するユーザ応答の認識結果も、他のユーザ発話の認識結果と同様に誤りの可能性がある。

自らの誤解を適切に検出・復旧できないシステムはタスクを遂行できないばかりか、システムの「独走」（図1）を招き、ユーザを不快にする可能性がある。「気軽さ」という音声対話の利点を損ねいためには、ユーザに特別な話し方や操作をさせることなく、あくまで自然な対話の中でシステムの誤解を克服していく技術が求められる。

2 関連研究

ユーザとの対話のやりとりの中から音声対話システムの誤解を検出・復旧しようとする研究は少ない。特に誤認識（聞き間違い）レベルの誤解は人間－人間の対話では生じにくく、人間－コンピュータの対話に特有の現象であるためこれまで十分に研究されてこなかった。従来、誤認識の訂正は、対話過程の中ではなく音声認識・理解研究の一部として行われてきた。すなわち、ユーザ入力の認識・理解結果が本来の入力と異なる（と思われる）時に、音声対話システムは、ユーザ発話の事例（書き起こしデータ）[1, 4]や書き起こしと認識結果の対応[8]などを用いて、認識結果を訂正するアプローチであった。

我々のアプローチはこれらの従来法とは異なり、誤認識の検出と復旧のための情報源として「ユーザとのやりとり（interaction）」を用いる。すなわち、システムがあらかじめ蓄えている知識から認識結果の誤りの検出と訂正を行うのではなく、「現在の認識結果に基づくシステム応答」とそれに対する反応である「次のユーザ発話」との関係を見ることで、対話が順調に進んでいるのか、それともやりとりに何らかの異常事態が発生しているのかを検出しようとするものである。

ユーザの音声入力を単体でとらえるのではなく「システムからの働きかけに対する反応」としてとらえる研究もいくつか存在する。たとえば[9, 7]は、システムの誤解に対するユーザの訂正応答（user correction）の発声には音響的な特徴があることを示している。また[2]はシステムの発する語彙が次のユーザ応答の語彙選択に影響を与えることを指摘している。これらから「次のユーザ応答」には「直前のシステム発話」で生じている事態が反映されていると考えられ、システムの理解状態の正誤の判定にも利用できるのではないかと思われる。

実際、ユーザの応答から直前のシステム発話の誤りを検出する試みも既に始められている。Levow[6]は最初のユーザ発話（original）とシステムの誤認識後

ユーザ	あの一月曜日の予約状況なんですけどぉ
システム	木曜日ノ予約ノ確認デスネ？
ユーザ	いや月曜日の予約なんですけど
システム	ハイ。木曜日ノ予約状況ハ...
：	：

図 1: システムが誤解を検出できない場合の対話

のユーザ発話（repeated user correction）の識別に関して、音響的パラメータだけによる決定木学習から75% の accuracy を得た。また Krahmer ら [5] はユーザ発話の形式的特徴を用いてふたつの確認方法（explicit と implicit）がシステムの誤解事態を検出する精度を比較報告している。本研究ではこれらとは異なり、人工的にシステムの誤解を生じさせ、それに対するユーザ応答を効率的に収集した。

3 実験

音声対話システムが誤解に基づいて発話する場合のユーザ応答の特徴を明らかにするため、対話収集実験を行った。「ユーザから音声が入力された後に、入力された内容をシステムが確認する」という場面を想定し、システムの確認内容が正しい場合と間違っている場合でユーザの応答がどのように異なるのかを明らかにすることが本実験の課題である。システムによる内容確認発話に対してユーザは「はい / いいえ」のような単純な答え方をするのか、もししないとすればどのように応答するのか、その応答は「正しい内容の受理」の場合と「間違った内容の非受理 / 訂正」の場合で異なるのか、などが実験で注目すべき点である。

もしシステムの確認に対するユーザ応答を、その特徴からふたつに分類することができれば、逆に、得られたユーザ応答を用いて、直前にシステムが確認した内容の正誤を判定できる見込みがある。すなわち、音声対話システムが対話のやりとりの中から自らの誤認識や誤解を検出することができる。本実験では、システムの発話内容が正しい場合と誤っている場合のユーザ応答を効率よく、適切な割合で収集するため、システムが対話の主導権を取り、wizard of OZ 法による実験を行った。

目的 音声対話システムが「正しく理解している」場合と「誤解している」場合を設定し、その理解状態に基づいてシステムがユーザに確認する時、システムの確認に対するユーザの応答が「正しい確認」と「誤解した確認」のそれぞれの事態においてどのように異なるか、その特徴の分析を目的とする。

実験環境 収録実験は、音声対話システム側の認識性能（誤認識が起こる程度）を統制するため wizard of OZ 法を採用した。従って、システムの誤解の仕方は予め実験計画で定めてあり、システムは被験者（ユーザ）の発声内容と関係なく、誤解を発生する。システムの音声には規則合成音声[1]を用い、被験者とシステムのやりとりは音声のみで行われた。

話し手	タイミング	発話例
システム	initQ	“何曜日デスカ？”
↓		(スロット名を prompt する質問)
ユーザ	応答 0	“月曜日です”
↓	(original)	
システム	確認 1	“火曜日デスカ？”
↓		(誤った確認発話)
ユーザ	応答 1	“違います。月曜日です”
↓	(response1)	
システム	確認 2	“木曜日デスカ？”
↓		(誤った確認発話)
ユーザ	応答 2	“月曜日です”
↓	(response2)	
システム	initQ	“何時カラデスカ？”
↓		(次のスロットへ)

図 2: 対話の構成（曜日スロットの例）

手続き 被験者は 10 ～ 40 代の大学生及び研究者、男女 13 名を用いた。被験者はまず「コンピュータと対話することで会議室の予約を取る」よう指示された。被験者ひとりあたり 5 ～ 10 対話を収録した。

収録した対話 対話のドメインは「会議室の予約」で、曜日・予約時間・会議室名の 3 つのスロットに値を入れる form-based の対話である。対話の主導権（initiative）は終始システム側にある。対話はシステムから話し始め、曜日→予約時間→会議室名の順にスロットを埋める値をユーザに尋ねる。すべてのスロットが埋まるとシステムが対話を終了させる。

各スロットでは図 2 に示すような対話が行われる。ひとつのスロットでのシステムの発話回数は、最小 1 回（initQ を尋ねるだけでユーザ応答 0 を受け取ると次のスロットに進む）から、最大 3 回（initQ と 2 回の確認を行ってから次のスロットに進む）の場合がある。

システムは確認 1・確認 2 のそれぞれの時点で、表 1 に示したシステム発話のタイプの中から実験計画に基づいてユーザに確認発話を行う。CR のシステム発話は、システムが正しい認識結果（Correct Recog.）を得た状態を想定しユーザが入力した正しい内容の確

[1] NTT サイバースペース研究所メディア処理プロジェクトのテキスト合成システム（FLUET）を使用した。

表 1: システム発話のタイプ

タイプ	発話の仕方 (相当する認識結果)	システム発話例
initQ	スロット名 prompt	“何曜日デスカ？” “ドノ会議室デスカ？”
CR	正しく確認する (Correct Recog.)	“月曜日デスカ？”
ME	間違えて確認する (Misrecog. Error)	“火曜日デスカ？”
ME2	値を変えて間違える (2度目の確認時のみ)	“木曜日デスカ？”
RE	質問しなおす (Rejection Error)	“何曜日デスカ？” “モウ一度オ願イシマス”

認を行う。ME は誤った認識結果（Misrecog. Error）を得た状態を想定し誤った確認を行う。RE は認識結果が得られなかった（Rejection Error）状態を想定しもう一度ユーザに質問しなおす。

4 結果

13 名の被験者により 120 対話（360 スロット分）が収録された。各スロットは 1 個〜3 個のシステム発話が含まれており、合計で 771 個のシステム発話（表 2）とそれに対するユーザの反応を収集した。

ユーザの無応答 771 のシステム発話に対して 769 のユーザ応答があった。すなわち、システムの発話に対してユーザが何も応答しない例が 2 回あった。この 2 例は同一の被験者で、応答がない直前のシステム発話のタイプはどちらも ME（間違えて確認する）だった。数は少ないが、システムの不適切な確認に対してユーザが「何を発声してよいかわからなくなる」場合があることを示唆している。

肯定 / 否定表現による応答 システムの発話のタイプ（表 1）のうち CR（正しい確認）と ME（間違えた確認）は yes/no 疑問なので、確認内容が正しければ「肯定」の応答が、確認内容が誤っていれば「否定」の応答が予期される。そこでユーザが実際に「肯定 / 否定」で応答するのかどうかを調べるため、ユーザ応答が肯定 / 否定を表わす表現を含む割合を調べた（表 3）。今回の収録では、肯定を表わす表現として「はい」「そうです」「そう」などがあり、否定を表わす表現として「いえ」「いいえ」「いや」「や」「違う」「違います」があった。

表 3を見ると、システムの正しい確認（CR）に対してはほぼすべてのユーザ応答（98.5%）に肯定表現

表 2: システム発話の組合せと頻度の一覧

システム発話のタイミング別			タイプ別
initQ	sys 確認 1	sys 確認 2	
initQ(360)	→なし (67)	→なし(67)	initQ(360) CR (131) ME (131) ME2 (18) RE (131)
	→CR (84)	→なし(64)	
		→CR(20)	
	→ME (98)	→なし (61)	
		→ME (19)	
		→ME2(18)	
	→RE(111)	→なし(50)	
		→CR (27)	
		→ME(14)	
		→RE (20)	
initQ(360)	sys1(293)	sys2(118)	計 (771)

が含まれているのに較べて、システムの間違った確認（ME と ME2）に対するユーザ応答が否定表現を含んでいる例は 4 割程度であった。さらに、否定表現だけで応答している例、すなわち「いいえ月曜日です」のように訂正を含む応答ではなく、「いいえ」や「違います」だけのユーザ応答はさらに少なく、システムの間違った確認（ME と ME2）に対するユーザ応答の 23.1% しかなかった。

この結果から、ユーザはシステムの正しい確認に対してほぼ必ず肯定表現で応答するのに対して、誤った確認に対しては必ずしも否定表現を使うとは限らない、ということが言える。従って、システムの確認発話の正誤をユーザ応答の肯定 / 否定表現から判定しようとする戦略は、確認内容が正しい場合には成立しても、確認内容が誤っている場合は得策とは言えないかもしれない。

内容語による応答 では、システムの誤った確認に対してユーザは否定表現を用いずに何を応答するのだろうか？多くの場合システムの誤った確認に対してユーザは訂正発話を行う。つまり、システムの確認内容が誤っているので（否定するのではなく）正しい内容を繰り返すことでシステムの誤解を訂正しようとする。そこで各種のシステム発話に対するユーザ応答中に、内容語（スロットを埋める表現。今回のタスクでは曜日・時間・部屋名）が含まれる割合を調べた（図 3）。

図 3を見るとシステムの誤った確認（ME・ME2）に対するユーザ応答には 7〜8 割前後の確率で内容語（i.e. 正しいスロット値の繰り返し）が含まれている。それに対してシステムが正しい確認（CR）をし

表 3: 肯定 / 否定の表現を含むユーザ応答の割合

タイミング	ユーザ応答 1	ユーザ応答 2			合計
前のシステム発話タイプ	CR	CR → CR	RE → CR		
肯定表現を含む応答の割合 (度数)	98.8% (83/84)	100% (20/20)	96.3% (26/27)		98.5% (129/131)
前のシステム発話タイプ	ME	ME → ME	ME → ME2	RE → ME	
否定表現を含む応答の割合 (度数)	38.1% (37/97)	63.2% (12/19)	38.9% (7/18)	38.5% (5/13)	41.5% (61/147)
否定表現のみの応答の割合 (度数)	21.7% (21/97)	36.8% (7/19)	11.1% (2/18)	30.8% (4/13)	23.1% (34/147)

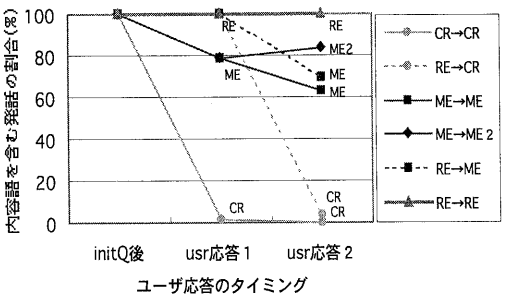

図 3: 内容語を含むユーザ応答の割合 (%)

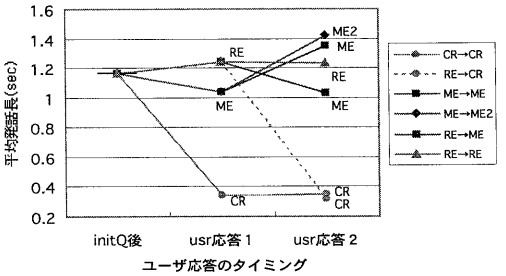

図 4: ユーザ応答の平均発話継続長 (sec.)

ている時はユーザはほとんど内容語を発話していない。なお、システムからの質問 (initQ) や質問の繰り返し (RE) は「何曜日ですか？」のような wh 疑問なので、それに対するユーザ応答は 100% 内容語を含んでいる。

誤った確認 (ME・ME2) に対するユーザの応答についてさらに詳しく見ると、システムが同じ誤りを 2 回繰り返す場合 (ME → ME) は 1 回目よりも 2 回目の誤りに対する時の方が内容語を含む発話の割合が下がり (78.4% → 63.2%)、否定表現の含まれる割合が上がっている (38.1% → 63.2%. 表 3)。それに対して、同じく 2 回間違える場合でも 1 回目と 2 回目でシステムが異なる誤りを犯す場合 (ME → ME2) は 2 回目の誤りに対する応答が内容語を含む割合は逆に高まる (78.4% → 83.3%)。システムが同じ間違いを繰り返すとユーザは訂正するより否定しようとするが、システムが異なる間違いをする時はユーザは否定するより訂正することに注意が向くのかもしれない。

以上のように、システムの確認発話に対するユーザ応答が内容語を含む割合を見てみると、システムの間違った確認に対するユーザの応答は概して内容語を用いて「訂正」を行う傾向が高いのに対して、システムが正しい確認をしている時のユーザ応答はほとんど内容語を繰り返さず肯定するだけだ、ということがわかる。

ユーザ応答の平均継続長　これまでユーザ応答の特徴を内容 (書き起こし) から分析してきた。しかし実際の音声対話システムではユーザの発話内容は音声認識器を経由して「認識結果」として得られるのであり、発話内容をそのまま得られる訳ではない。そこで認識器 (音響モデル・言語モデル) の性能に依存しない音響的な特徴量として、ユーザ応答の平均継続長による分析を行った (図 4)。

図 4 を見ると、最初のシステムからの質問 (initQ) に対するユーザ応答の平均発話継続長が 1.16sec であるのに比べると、システムの正しい確認に対するユーザ応答 (肯定応答) はとても短い応答 (0.3sec 程度) になっている。もちろんこの差はユーザが応答すべき内容語の単語長に依存するものであるが、システムからの再質問 (RE) や間違った確認 (ME) に対するユーザ応答の発話長と比べると正しい確認 (CR) に対するユーザ応答の発話長は、明らかに短い傾向がある。これは前述したように、正しい確認に対するユーザの応答は、内容語を含まず肯定表現のみ (「はい」

や「はいそうです」など）でなされることが多いためと考えられる。

また、興味深いのは、システムの確認が2回間違う場合（ME→ME、ME→ME2）のユーザ応答は、1回目に間違う場合の応答に比べて長くなっている。これは2回目のシステムの誤りにはユーザの応答が強調されるためかもしれず、他の音響的特徴（ピッチやパワー）も併せた分析が必要であろう。

5 考察

以上のように、システムが正しい確認（CR）をしている時、それに対するユーザ応答は肯定表現が多用されほとんど内容語を含まず、従って発話長は短くなる。それに対して、システムが間違った確認（MEとME2）を行う場合は、ユーザ応答には必ずしも否定表現は使われず、正しい内容語を繰り返すことで訂正が行われ、従って発話長は長めになる。

ユーザと音声対話システムのやりとりが「コミュニケーション」であることを考えれば、これらの結果は至極当然である。相手が誤解している時に「違う」と言うだけではタスクは遂行されず、「どう違うか」つまり「訂正」する方が協調的である。また、相手が誤解していない時にはコミュニケーションは順調に進んでいるのであり、正しい確認に対しては必要最低限のことを言うだけでよい。わかりきったことを冗長に繰り返すのはコミュニケーションの効率を下げるので、この時のユーザの反応も理にかなっていると言えるだろう。

今回の実験では、wizard of OZ法を採用したのでシステムの誤解に対するユーザの反応に関して基本的なデータを網羅的に集めることができた。しかしながら、実験上、完全にシステム主導の対話とせざるをえず、ともするとユーザは固く、かしこまり、システムに聞かれたことだけに応答する、という印象の対話になった。今後は、ユーザが自由に対話できるよう、より主導権が交代しやすい対話の中で、音声対話システムが実際の音声認識結果を用いて、ユーザとの対話のやりとりの中から自らの誤解を検出・復旧する実験を行う予定である。

6 おわりに

音声対話システムがユーザとの対話のやりとりの中から、システム自身の誤解を検出・復旧するため、まずシステムが誤解している場合のユーザ応答の特徴を分析した。システムの誤解に対するユーザの応答は、必ずしも否定表現が含まれず、訂正を行うため内容語が繰り返され、発話長も長めになることがわかった。システムが正しい時のユーザの応答と、システムが誤解している時のユーザの応答の間には明白な傾向の違いがあり、システムが対話の中から自らの誤りを検出できる可能性が示された。

謝辞 日頃よりご指導いただく、NTTコミュニケーション科学基礎研究所 メディア情報研究部 萩田紀博部長、有益な示唆をいただく対話研究グループの諸氏、実験の準備にご助言いただいた木間良子さん、分析にご協力いただいた慶應大学の酒巻隆治さん、実験の被験者のみなさまに感謝いたします。

参考文献

[1] T.-H. Chiang and Y.-C. Lin. Error recovery for robust language understanding in spoken dialogue systems. In *Eurospeech99*, pp. 2007–2010, 1999.

[2] J. Gustafson, A. Larsson, R. Carlson, and K. Hellman. How do system questions influence lexical choices in user answeres? In *Eurospeech97*, 1997.

[3] B. A. Hockey, D. Rossen-Knill, B. Spejewski, M. Stone, and S. Isard. Can you predict responses to yes/no questions? yes,no, and stuff. In *Eurospeech97*, 1997.

[4] 石川開, 隅田英一郎. テキストコーパスを用いた音声認識誤り訂正手法. 言語処理学会 第5回年次大会 発表論文集, pp. 100–103, 1999.

[5] E. Krahmer, M. Swerts, M. Theune, and M. Weegels. Problem spotting in human-machine interaction. In *Eurospeech99*, volume 3, pp. 1423–1426, 1999.

[6] G.-A. Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. In *COLING-ACL98*, 1998.

[7] S. Oviatt, G.-A. Levow, M. MacEachern, and K. Kuhn. Modeling hyperarticulate speech during human-computer error resolution. In *ICSLP96*, 1996.

[8] E. K. Ringger and J. F. Allen. A fertility channel model for post-correction of continuous speech recognition. In *ICSLP96*, 1996.

[9] 宇津木成介, 竹内由則. 音声認識装置が音声を正しく認識しない事態における人間の発声の変化. 人間工学, 31(4):287–293, 1995.