音声対話システムの誤解状態におけるユーザ応答の分析

菊池　英明† 林　美穂子‡ 青山　一美‡ 白井　克彦‡

kikuchi@kokken.go.jp {mihoko, kazumi, ks}@shirai.info.waseda.ac.jp

†国立国語研究所 ‡早稲田大学理工学部

1. はじめに

人間同士の対話では、対話の進行とともに対話参加者の感情や心理状態が様々に変化し、多かれ少なかれ互いに相手の状態を認識しながら対話を行っている。従来の音声対話システムにおいては、相手の表情、声の調子、しぐさなどを判断して応答を柔軟に変化させる機能を持たないために、ユーザの状態に関わらず一定の調子で対応をするだけであった。本研究では、ユーザの感情や心理状態を推測しながら状況に適した行為を振る舞うことが可能な音声対話システムの構築を目指している。

ここで、音声対話システムと人間のコミュニケーションにおいて、現状では音声の誤認識・理解は避ける事ができない大きな問題である。この問題に対して、韻律情報を用いてシステムの誤認識を検出する方法[1]、発話継続長を中心とした複数のパラメータにより総合的にシステムの誤認識を検出する方法[2]が有効であることが示されている。本研究ではさらに円滑なコミュニケーションを実現するためにさらに、システムが自分自身の誤解を認識したうえで、誤解が引き起こしたユーザの状態の変化を推測することで、状況にふさわしい対応をとることが重要と考える。例えば、ユーザが初めて経験するシステムの誤解に対して困惑におちいった場合に誤解について丁寧に説明したり、繰り返されるシステムの誤解に対してユーザが怒りの感情を表した場合に誤解が一切なくなるように対話戦略を変更したりするなどが考えられる。

そこで本研究では、まずシステムの誤解状態を意図的に発生させた模擬対話を収録し、対話中でのシステムの確認発話に対するユーザ応答分析を行い、ユーザの感情や心理状態の認識に役立つ特徴が得られるかを検証した。本稿では特に言語情報、応答の早さ、パワー、ピッチについての分析結果について報告する。

2. 模擬対話収録実験

2.1. 収録環境

システムの誤解に対するユーザ応答の特徴を調べるために、一定の条件で誤解を発生させる必要がある。従って Wizard of OZ 方式で模擬対話収録実験を行った。被験者は独立したブース内でヘッドフォンからシステム応答（合成音）を聞き、ヘッドセットマイクを使って発話する。別のブースで操作者がシステムの誤解発生状況を制御する。なお、誤解発生状況としては、被験者にとって誤解の認識が容易と考えられるシステムからの確認発話に限定した。操作者は、確認方法（直接・間接）と誤解の有無を制御する。

収録した対話のドメインは「チケット予約」で、対話の終了条件は「公演名・公演会場名・公演日・公演時間・予約枚数の５つのスロットを埋めてチケットの予約を完了すること」とした。システム側が主導権をもち、ユーザの発話を促す形で対話を進める。被験者にはあらかじめチケット予約一覧を渡し、コンピュータと対話してこの一覧に記載されているチケットを予約するよう指示した。

2.2. 収録対話の概要

被験者一人あたり 20～30 対話、学生 4 名、他 4 名の計 8 名の被験者により、合計 196 対話、211.1 分を収集、確認発話に関するシステムとユーザのやりとり 1310 個を収集した。

対話音声を DAT に収録し計算機に取り込んだうえで、転記ツールを用いて書き起こし作業を行った。さらにシステムの確認発話に対するユーザの応答発話にタグを付与した。これらのデータを用いて、発話内容、ポーズ長、パワー、ピッチなどの分析を行った。図１に収録した模擬対話の例、表１に被験者応答の概要を示す。

S:いらっしゃいませ、チケットを予約します。
S:ご希望の公演名をお願いします。
U:吉田拓郎をお願いします。
S:吉田拓郎ですね。
U:はい、そうです。[Explicit:Right:Yes]
S:ご希望の公演会場はどこですか？
U:大宮ソニックシティです。
S:東京国際フォーラムホール A ですね。
U:いいえ、大宮ソニックシティです。
[Explicit:Miss:Correct]
S:大宮ソニックシティ大ホールですね。
U:はい、そうです。[Explicit2:Right:Yes]
S:ご希望公演日をお願いします。
U:8 月 10 日です。
S:8 月 9 日ですね。
U:いいえ、違います。[Explicit:Miss:No]
S:ご希望公演日をお願いします。
…
(以下省略)

図 1：収録対話例

(S はシステム、U は被験者の発話を示す。
[]内はそれぞれの発話に付与されたタグを示す。)

表 1：収集された被験者応答の概要

(()内は連続して行った 2 回目の確認を除いた数)

総発話数		2251
確認応答数		1310
直接確認	正解	376(234)
	誤解	371(226)
間接確認	正解	267(242)
	誤解	296(247)

以下には収集した被験者応答について、言語情報、パラ言語情報の特徴を分析した結果を示す。

3. 言語情報の分析

ここでは、システムの確認発話に対する被験者の応答に対して確認および応答の種類をタグ付けした結果に基づいて分析を行う。

3.1. 確認の正誤による違い

表 2 に確認発話に対する被験者応答の分類とその割合を示す。なお、分析にはシステムが連続して確認を行った場合の 2 回目を除いた（表 1 参照）応答を対象とした。

表 2：確認発話に対する被験者応答の分類

システム確認		被験者応答	総数（割合[%]）
直接確認	正解	肯定のみ	234(100.00)
	誤解	否定のみ	127(56.19)
		否定訂正	7(3.10)
		訂正のみ	90(39.82)
		その他	2(0.89)
間接確認	正解	次の値	233(96.28)
		その他	9(3.72)
	誤解	否定のみ	45(18.22)
		否定訂正	14(5.67)
		訂正のみ	79(31.98)
		訂正+次	10(4.05)
		次の値	99(40.08)

直接確認では、正解と誤解で同じ種類の応答が現れることはなかった。つまり、応答の種類（肯定、否定、訂正など）が正しく分析できればシステムの誤解状態は完全に検出できることになる。

一方、間接確認においては、正解だけでなく誤解状態においても次のやりとりに進む傾向が強く（表中の「次の値」）、これは主に実験における被験者の動機付けの失敗に起因すると考えられる。しかしながら間接確認の方が誤解の解消を行わないことへの抵抗が小さいことも予想でき、したがって早期の誤解検出のためには言語情報以外の情報からこうした状態の検出を行うことが重要といえる。

3.2. 対話回数による変化

さらに、誤った確認に対する同一被験者の応答を対話回数に着目して分析した。

直接確認は yes/no 疑問であるため、終始「いいえ」や「違います」などの否定のみを行う被験者と、始めから訂正を行う被験者に分類された。また、数名の被験者には徐々に訂正発話が増えて行くという傾向が見られた。

間接確認では、システムの誤解を指摘するか、次のスロットに進むかの 2 通りがあり、必ずしもシステムの誤解を指摘しているわけではない。しかし、始めはシステムの誤解を指摘せずに次のスロットに進行していたが、対話を繰り返すうちに、徐々に訂正を行うようになる被験者が多かった（図 2 参照）。特に顕著な例を図 3 に示す。

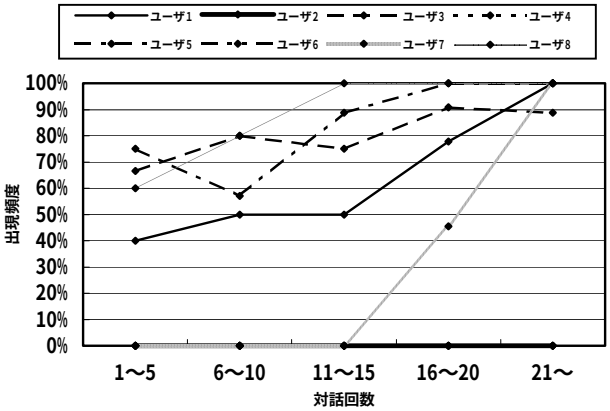

図２：間接確認における訂正発話の出現頻度の推移

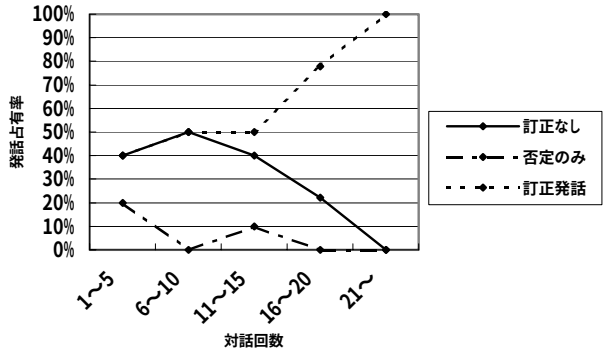

図３：間接確認に対する応答方法の変化（被験者７）

3.3.　考察

システムの確認発話に対する被験者応答の発話内容を分析した結果、システムの正しい確認に対する応答は肯定のみのため、対話を繰り返しても大きな変化は見られなかったが、システムの誤った確認に対しての応答にはいくつかの傾向が見られた。１つは、否定のみの発話より訂正発話が多く行われていることである。特に今回の実験では、被験者がシステムの誤解を指摘すると、システムは同じ質問を繰り返した。そのため、被験者は自然にシステムの次の発話を予測して訂正発話を行い、繰り返される質問を避けて効率を上げようと試みているのではないかとも考えられる。

また、対話回数の増加に伴う個々の被験者の対応は様々であり、慣れていくスピードにも差が見られたが、最終的にはほとんどの被験者がシステムの動作や発話に慣れ、効率のよい対応を取るようになった。この結果からも、システムの振る舞いに対するユーザの慣れとともに変化する応答方法からユーザの状態を的確に推測できれば、より円滑に対話が進められるのではないかと考えられる。

4.　パラ言語情報の分析

ユーザの困惑状態や怒りの状態の検出に役立つ特徴を得るために、ポーズ長、パワー、ピッチの分析を行った。

4.1. 応答の早さ

被験者の応答の早さとしてシステムの確認発話終了から被験者の応答開始までのポーズ長を計測した。

4.1.1. 確認の正誤による違い

確認方法およびその正誤とポーズ長平均値の関係を表３に示す。

全体を通して、誤解時のポーズ長は正解時と比較して長くなっている。直接確認において、システムが正しく認識した場合、ユーザは肯定するのみであるため、あまり思考時間を必要としない。それに対して、システムが誤った認識をした場合、ユーザは次にどのような発話をすればいいのか困惑し、発話するまでにある程度の思考時間を必要としていると考えられる。間接確認において差が現れなかったのは、システムの認識が正しい場合でも次のスロットの解答を発話するのにある程度の思考時間を必要とするためと考えられる。

表３：確認方法およびその正誤とポーズ長

システム確認の分類		平均ポーズ長[msec]
直接確認	正解	269.67
	誤解	540.04
間接確認	正解	690.63
	誤解	705.51
全体	正解	440.35
	誤解	612.00

4.1.2　対話回数による変化

3.2 と同様に誤った確認に対する同一被験者の応答を対話回数に着目して分析した。特に変化が著しかった間接確認に対する訂正発話のケースについて対話回数増加に伴うポーズ長の推移を図４に示す。この図に見られるように、大局的に見れば対話回数の増加に伴って被験者の応答は早くなっていく。これはシステムの振る舞いに対するユーザの慣れによるものと考えられる。

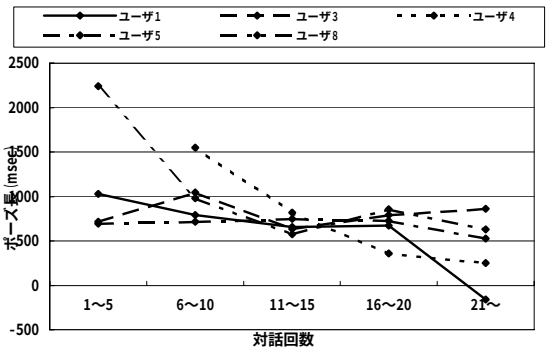

図４：誤った間接確認に対する訂正発話のポーズ長の対話回数増加に伴う変化

4.1.3. 対話の進行状況による違い

同一被験者の一対話中のポーズ長の変化を見た。具体的には、対話の進行状況が顕著に異なる例として五つのスロットにおける確認発話が全て正しかった場合と、全て誤った場合の二つを比較した。その結果を図５、図６に示す。

この結果から、正しい確認が続いている間は、徐々にシステムの確認発話に対するユーザのポーズ長は短くなっていく傾向が得られた。しかし、誤った確認が行われていた場合、ユーザのポーズ長にばらつきが見られる。

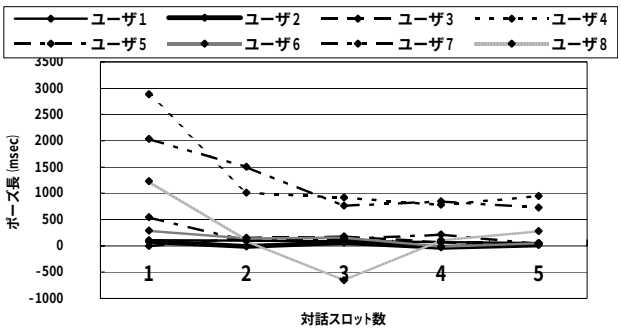

図５：全て正しい確認発話に対するポーズ長の変化

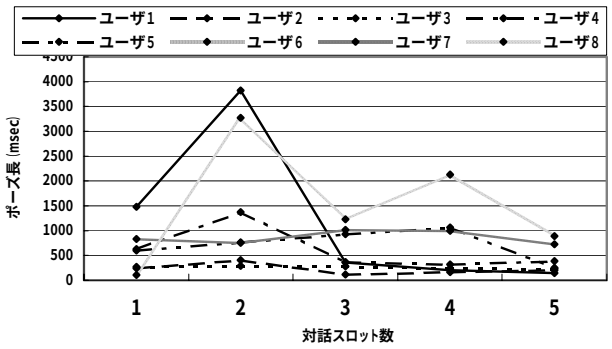

図６：全て誤った確認発話に対するポーズ長の変化

さらに、これまでの分析結果から、ユーザにとって初めてシステムが誤解した場合に心理状態は大きく変化し、ポーズ長にその影響が現れるのではないかと考え、分析を行った。その結果を図７に示す。

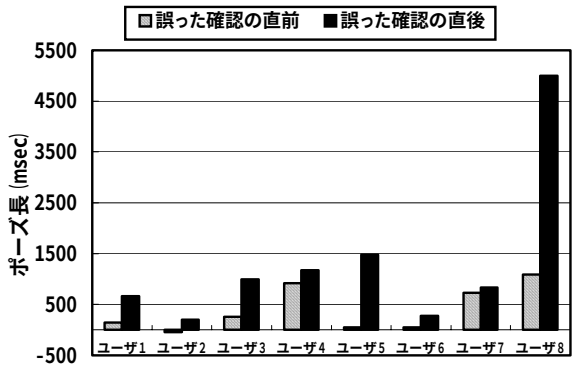

図７：システムが最初に誤った前後のポーズ長変化

図７から、全ての被験者についてシステムの最初の誤った確認に対する応答が遅れていることがわかる。以上の結果から、システムが正しい認識を行ううちは対話のテンポも徐々によくなりユーザの応答も早くなっていくが、一般的にシステムが誤った認識を行うと応答に遅れが生じるといえる。特にユーザにとって初めてシステムの誤解が生じた時、ユーザが困惑状態に陥り極端にポーズ長が長くなる傾向があるといえる。

4.2 応答のパワー

400[msec]以上の無音区間で区切られる音声区間を発話単位とし、発話単位毎に短時間平均パワーを計測した。

4.2.1. 確認の正誤による違い

システムの誤った確認に対する応答において、何らかの強調を行うことが予想される。そこで、システムの認識状態、確認方法、被験者の応答方法別にパワーの分析を行った。その結果を表４に示す。

この結果から、システムが正しい確認を行った場合よりも、誤った確認を行った場合の方が被験者応答のパワーが大きいことがわかる。また、直接確認よりも間接確認に対する被験者応答のパワーが大きい。被験者ごとに見ても、特に訂正発話において、その傾向が強く見られた。間接確認では、システムが誤った認識をしていても、次のスロットに進んでしまうので、それを阻止するために強調しているのではないかと考えられる。

表４：確認方法およびその正誤と被験者応答のパワー（最大値）

システム確認		被験者応答	全被験者の平均
直接確認	正解	肯定	4.55E+07
	誤解	否定のみ	6.85E+07
		訂正発話	6.51E+07
間接確認	正解	肯定	7.73E+07
	誤解	否定のみ	1.28E+08
		訂正発話	8.11E+07
全体	正解	肯定	5.90E+07
	誤解	否定のみ	8.27E+07
		訂正発話	7.24E+07

4.2.2. 対話の進行状況による違い

4.1.3 と同じように、システムの誤りに初めて直面した場合、応答の遅れだけではなく声の調子にも変化が出るのではないかと考え、分析を行った。

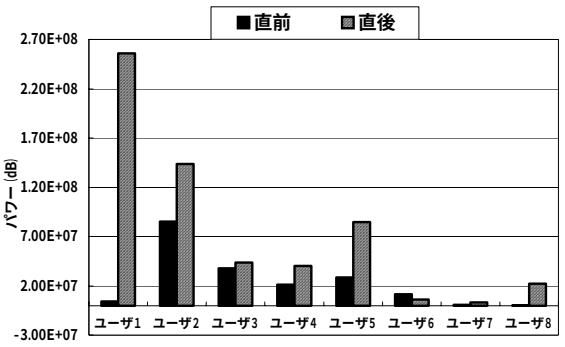

図８：システムの最初の誤解の前後のパワー変化

図８を見ると、８人中７人の被験者のパワーが増加していることがわかる。システムの誤解を指摘し、自分が伝えたい情報が間違って伝わってしまうことを避けるために、否定や訂正を強い調子で行うためと考えられる。

以上から、システムの誤りに対して、ユーザがシステムの誤解を指摘して正しい内容を伝達しようとする、あるいは対話の進行を止めるために、強調の作用が働き、その結果パワーが通常より大きくなると考えられる。なお、実験において繰り返されるシステムの誤りに対して、被験者の中には明らかに怒りの感情を現しているケースもあった。こうした感情の変化がパワー値から認識されることが期待されるが、そのためには対話の進行状況による影響を考慮する必要がある。

4.3 ピッチ分析

怒りを含む感情の判別において話者毎に異なるものの F0 の最大値、変動範囲の利用が有効であることが報告されている[3]。ここでは発話単位毎に F0 を抽出してスムージングを行った結果から F0 の最大値、平均値などを求めた。

4.3.1. 確認の正誤による違い

表５に、確認方法およびその正誤の違いに対する F0 平均値と最大値の平均を示す。

表５：確認方法およびその正誤とユーザ応答のピッチ

システムの状態			F0 平均値 [Hz]	F0 最大値 [Hz]
直接確認	正解	肯定	183.3	207.6
	誤解	否定のみ	175.0	205.2
		訂正発話	187.3	248.7
間接確認	正解	肯定	176.5	233.5
	誤解	否定のみ	173.6	239.5
		訂正発話	194.4	255.1
全体	正解	肯定	178.3	217.5
	誤解	否定のみ	179.5	230.0
		訂正発話	186.4	252.3

この表において直接確認より間接確認の場合の方がピッチが高くなる傾向が見られる。また、誤解に対して訂正発話で応答する場合が最も平均値・最大値ともに大きくなる傾向が見られた。

4.3.2. 対話の進行状況による変化

システムの初めての誤解に対して、ポーズ長やパワーと同様に比較分析を行った結果を図９に示す。

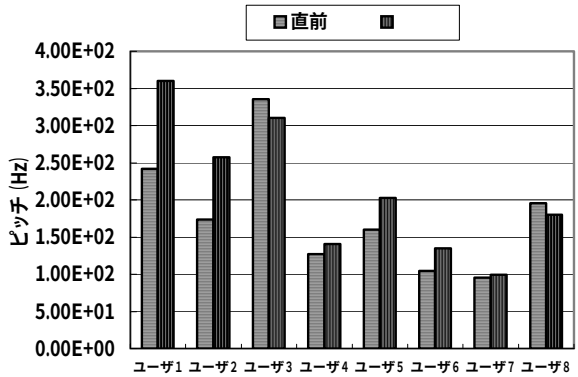

図９：システムの最初の誤解の前後のピッチ（最大値）の変化

図９において、８人中６人の被験者はシステムの誤った確認に対して F0 最大値が増加している。このように全体の傾向として誤解に対してピッチが大きくなる傾向が見られるが、その傾向が全く見られない被験者も存在する。こうしたパラ言語情報の表出においては話者依存性が大きいことが予想される。

4.4. 考察

システムの確認発話に対する被験者応答のパラ言語情報を分析した結果、確認の種類やその正誤による変化の他、対話回数、対話進行状況による各特徴量の変化が見られた。典型的には、対話回数の増加に伴いシステムの振る舞いに対する慣れが増すためか応答が早くなる他、システムの初めての誤解に対しては戸惑いのためかパワーやピッチの増加がより大きくなる、などの傾向が見られた。こうした慣れや戸惑いなどのユーザの状態は、話者による量的な違いはあってもパラ言語情報として表出されており、システムがこれを抽出して利用するに値するといえる。

5. おわりに

本稿では、システムの誤解状態を意図的に発生させた模擬対話において、被験者の応答の種類、応答の早さ、応答発話のパワー・ピッチなどについて分析した結果について述べた。応答の種類としては、対話回数の増加に伴い訂正発話が増えていく傾向が見られた。応答の早さについて、全般的に誤解時に遅れが生じる傾向があるが、特に初めての誤解の際にその傾向が顕著にあらわれることがわかった。また応答発話のパワーおよびピッチについても、初めての誤解の際に最大値が通常より大きくなる傾向が見られた。

今後、これらの分析結果に基づいて、対話を円滑に進めるための対話戦略制御とそれに必要なパラメータの関係を明らかにした上でモデル化を行い、これまでに開発してきた汎用的プラットフォーム[4]に導入する予定である。

参考文献

[1] G.A. Levow, “Characterizing and Recognizing Spoken Corrections in Human-Computer Dialogue,” Proc. COLING-ACL98, 1998.

[2] 平沢, 宮崎, 相川, “質問-応答連鎖からの音声対話システムの誤解の検出,” 情報処理学会研究報告, SLP-34, pp.239-244, 2001.

[3] 門谷, 阿曽, 鈴木, 牧野, “音声に含まれる感情の判別に関する検討,” 情報処理学会研究報告, SLP-34, pp.43-48, 2001.

[4] 青山, 平野, 菊池, 坪川, 白井, "音声対話システム汎用プラットフォームの検討, 情報処理学会研究報告, SLP-30, pp.7-12, 2000.