音声対話システムの相槌応答タイミングによるユーザの印象への効果

平沢純一　中野幹生　川端豪

ＮＴＴ研究所

あらまし 実時間の音声対話システムを設計するには，システムの応答内容だけでなく，応答タイミングを考慮することが重要である．人間同士の対話では，相槌を始めとする聞き手の応答タイミングは話し手の発話に頻繁に重なる．しかし，対話システムの応答タイミングを，人間の応答と同様に話し手の発話に重ねることが，ユーザの快適さに貢献するかどうかは実証されていない．本研究ではWizard of OZ法を用いた対話実験により，システムの相槌応答タイミングをユーザ発話に重ねることがユーザの印象に与える効果を調べた．実験の結果，システム相槌の重なりは，ユーザのタイプに応じて正反対の印象を与えることが明らかになった．

1　はじめに

人間が快適に違和感なく計算機と情報交換できる音声対話インタフェースの実現を目指している．人間と音声対話システムが情報交換するには，システムが応答などを通じてシステムの内部状態を人間（ユーザ）に適切に開示する [3] ことが重要である．実時間システムの応答を適切にデザインするには，「何を」応答するかだけでなく，「いつ」応答するか（**応答タイミング**）も検討されなければならない [4]．システムの応答タイミングが不適切だと，ユーザを不快にするばかりでなく，システム応答の意図がユーザに誤解されたりタスク遂行を妨げる原因になる．

適切な応答タイミングは単純な話者交替 (orderly turn-taking) では必ずしもモデル化できない．人間同士で行われている対話は，ひとつのボールを交互に投げ合うように単純ではなく，聞き手の応答は相手発話の終りを待たずに重なって始められる [10, 7]．その割合は「対話中の45%の発話が相手から何らかのオーバーラップを受けている」との報告 [6] がある．

しかしながらこれらは人間同士の対話に関する知見であって，人間 - システムの対話において，工学的デザインの立場からシステム応答がユーザ発話に重なることの効果はまだ実証されていない．そこで本発表では，システムの相槌応答の適切なタイミングについて調べる．相槌を調べるのは，相槌はユーザの入力に対する最も基本的なシステム応答のひとつであり，対話においても重要な役割を果たすからである．本発表では，ユーザから入力された情報をシステムが受理したことを**相槌応答**で開示し，その相槌（開示）のタイミングとして (a) ユーザ発話の途中でも情報を受理した時点，(b) ユーザ発話に重ならずに発話末まで待った時点，の2条件を設定し，システム相槌タイミングの違いがユーザの印象に与える影響を調べることを目的とする．

2　従来の研究

人間 - 人間の対話を対象とした相槌応答の出力タイミングの研究は多い．それらは2つの立場に大別できる．すなわち，相槌タイミングをモデル化するための情報源として (1) 非言語情報（ピッチパタンなど）のみを利用する立場，(2) 言語情報（品詞やキーワードなど）まで含めて利用する立場，である[1]．

(1) の立場を支持する研究として綿貫ら [12] は，言語情報を取り除きピッチや映像だけを提示された被験者が相槌を挿入する実験を行い，被験者間での相槌挿入位置の相関が高いことから「人間は非言語情報を利用してあいづちを挿入している」と主張している．しかし綿貫らは「相槌生成に非言語情報が利用できる」ことを実証したに過ぎず，我々はここから「非言語情報**だけですべての**相槌生成をモデル化できる（相槌モデルに言語情報は不要）」と結論づけるべきでない．なぜなら (1) の立場は相槌の機能をひとつに限定してしまっている．

堀口 [5] は相槌の機能を5つに分類している．ひとつは単に「聞いているという信号」としての相槌である．非言語情報だけから相槌タイミングをモデル化する (1) の立場は，この「聞いている相槌」だけを対象にしてい

[1] 人間 - システム対話研究ではこの立場の違いは，システム相槌生成に音声認識を必要としない / する の違いとなる．

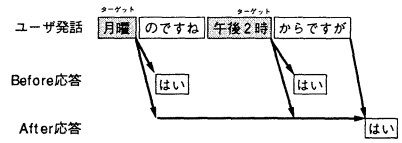

図 1: **実験条件 (Before 応答条件と After 応答条件)**

図 2: **Wizard of Oz 法による実験環境**

ると言える。しかし，非言語情報だけからでは，それ以外の「理解しているという信号」としての相槌などを生成することはできない．実際の音声対話システムにおいても「聞いている相槌」だけでユーザと意思疎通することは難しく，「理解している相槌」を含めた応答が必要となる．

しかしながら，人間 - システム対話の研究で，言語情報まで用いた「理解している相槌」の研究は十分ではない．「聞いている相槌」の研究としては，Ward[11] が非言語情報（韻律）だけを情報源とする相槌生成ルールを考案し，ユーザに機械だと気付かせない相槌生成システムを作った．西ら [8] は非言語情報による「聞いている相槌」を実現するシステムとして，ユーザ発話の音声区間の終了を検知すると相槌を返す音声メールシステムを作った．ユーザがこのシステムの好ましさを実験で評価したところ，意外なことに，システム相槌の有無では有意な差がなかった[2]．

これらに対して岡登ら [9] は，ユーザ発話にキーワードが含まれている場合にユーザ発話末で相槌を生成する実験システムにより，言語情報（発話内容）も用いた「理解している相槌」の有無の影響を調べた．実験の結果，システム相槌はユーザの発話を促進することが示された[3]．人間 - システムの対話で言語情報を用いた「理解している相槌」の効果を調べている研究は少なく，相槌タイミングまで含めた検討はされていない．

3 実験

3.1 目的

相槌応答タイミングの異なる 2 つの対話システムに対して，被験者が感じる印象の違いを調べることを目的とする．相槌には「システムが情報を受理（理解）したことを開示する」機能を担わせ，その開示のタイミングが（人間同士の対話で見られるように）相手発話に重なることの有効性を調べる．実験条件は以下の 2 つとする（図 1参照）．

[2]西ら [8] はまた，システム相槌がユーザ発話に重なるとユーザの評価が低かったと報告しているが詳細は述べられていない．

[3]ユーザ発話中のユーザ相槌の含有率が高まったことを以って，発話が促進された（≒人間同士の対話に近づいた）としている．

Before 応答条件: 情報（ターゲット表現）を受理した時点で，ユーザ発話の終了前でも (ユーザ発話に重なっても) 相槌応答を生成

After 応答条件: 受理した情報（ターゲット表現）を含むユーザ発話の終了後に (ユーザ発話に重ならずに) 相槌応答を生成

3.2 実験環境と装置

対話のドメインは「会議室の空き状況調べ」で，スロットフィリングタスクの構造 [2] を持つ．実験では音声認識の誤りに伴う要因を排除するため，Wizard of OZ 法（以下 WOZ 法）[1] を採用した（図 2）．

システムはユーザとの入出力とも音声のみとした．システムの音声には規則合成音声を用いず，女性の声をキャプチャした録音編集音声 (pre-recorded voice) を用いた．ユーザ発話中のシステム相槌を抑制する（After 応答条件）ため音声区間検出器を用いた．検出器は入力音声中のポーズ長（400ms），最小発話継続長（100ms）を閾値として音声区間を検出した．

実験者（wizard）は，(1) ユーザ音声を認識してターゲット表現を検出する，(2) 応答生成ルール（表 1）に従って応答生成トリガーを出す，だけでユーザ発話の意図認識などは行わない．この WOZ システムは，認識誤りのない音声認識器 (wizard) 以外は，現状の技術ですべて実装可能である．

3.3 実験方法

被験者は音声対話システム研究に関わりのない，20

表 1: **応答生成ルール**

ターゲット表現	トリガーのきっかけ	システム応答
スロットを埋める単語 ・月曜 (日) ・午前 9 時 ・第 1(会議室)	ターゲット表現の検出	相槌「はい」
あいさつや導入表現 ・すいません ・〜したい (んですが) ・〜をお願い (します)	ターゲット表現の検出	相槌「はい」
	全スロットの充足	確認発話 「えーと〜曜日の 〜時から〜時まで 第〜会議室ですか？」
	確認発話に対するユーザの了解表現	情報提供発話 「あいてますけど」 「予約が入ってますが」

表 2: 印象評定の平均値（システム相槌の重なりを意識しない群 / 敏感群）

被験者群	意識しない群（6 人 29 対話）				敏感群（4 人 25 対話）			
応答条件	Before 条件 (18 対話)	After 条件 (11 対話)			Before 条件 (15 対話)	After 条件 (10 対話)		
評価項目	平均 (S.D.)	平均 (S.D.)	t 値	有意水準	平均 (S.D.)	平均 (S.D.)	t 値	有意水準
使いやすい	*3.89*(0.57)	3.27(0.75)	t=2.42	$p < .05$	3.00(1.10)	*3.70*(0.49)	t=2.12	$p < .05$
話しやすい	3.67(0.67)	*3.82*(0.58)	t=0.60	n.s.	2.73(1.48)	*4.00*(0.78)	t=2.38	$p < .05$
快適だ	*3.83*(0.69)	3.73(0.62)	t=0.40	n.s.	2.73(0.93)	*3.40*(0.49)	t=2.00	$p < .10$
好き	*3.61*(0.59)	3.46(0.66)	t=0.64	n.s.	2.80(0.65)	*3.50*(0.50)	t=2.76	$p < .02$
親しみがある	*3.89*(0.31)	3.64(0.48)	t=1.65	$p < .20$	3.27(0.68)	*3.50*(0.81)	t=0.75	n.s.
人間的	*3.39*(0.83)	2.73(1.14)	t=1.75	$p < .10$	3.07(1.18)	*3.80*(0.98)	t=1.56	$p < .20$
わかりやすい	4.33(0.47)	*4.36*(0.64)	t=0.14	n.s.	3.73(0.68)	*3.90*(0.54)	t=0.62	n.s.
すばやい	*4.11*(0.81)	3.82(0.94)	t=0.86	n.s.	*3.93*(0.85)	3.10(0.54)	t=2.63	$p < .02$

S.D.= Standard Deviation(標準偏差)
n.s.= no significant(有意差なし)

発話番号	開始 -	終了	発話内容	フェーズ
S0	0.00 -	0.30	beep	(対話開始)
U1	1.20 -	2.79	会議室を予約したいんですが	[情報入力]
S2	3.04 -	3.34	はい	↓
U3	4.09 -	5.34	えーっと木曜日の	↓
S4	5.61 -	5.91	はい	↓
U5	6.09 -	7.77	午後 1 時から午後 4 時まで	↓
S6	7.95 -	8.25	はい	↓
U7	8.37 -	10.23	第 3 会議室をお願いしたいんですが	↓
S8	10.49 -	10.79	はい	↓
S9	10.80 -		えーと木曜日午後 1 時から	[確認]
		- 16.38	午後 4 時まで第 3 会議室ですか？	↓
U10	16.54 -	17.26	はいそうです	↓
S11	17.37 -	19.28	えーと空いてますけど	[情報提供]
U12	20.44 -	21.93	あそうですかわかりました	↓
S13	22.42 -	23.44	ありがとうございました	↓
S	23.97 -	24.39	beep	(対話終了)

U はユーザ（被験者）発話，S はシステム発話
発話開始時刻，終了時刻の単位は sec.

図 3: 対話例（After 応答条件）

～30 代の研究者と研究補助業務 10 人（男女 5 名づつ）とした．被験者はまず「コンピュータと対話をすることで会議室の予約状況を調べてもらいます」と教示され，調べる会議室の条件（曜日・予約開始時刻・終了時刻・会議室名）を与えられて対話を開始した．

1 対話を終了する度に被験者は 8 項目に関する 5 段階評定[4]を行った．1 人 8 対話を収録した後「WOZ 方式に気付いていないか」「実験条件（相槌タイミング）の違いに気付いたか」をインタビューで調べた．

4 実験結果と考察

被験者 10 人で全 80 対話を収録したうち，相槌タイミングが実験条件を充たしている 54 対話（Before 条件 33，After 条件 21）を分析対象とした．収録された対話の例（After 応答条件）を図 3に示す．システムによる平均相槌数は After 条件で 3.95，Before 条件で 4.97 だった．この差は，After 条件では 1 つのユーザ発話中に複数のターゲット表現が含まれる場合，まとめて 1 回のシステム相槌を生成するためである．実験条件による違いは対話前半（情報入力フェーズ）のシステム相槌タイミングと回数だけで，対話後半（確認フェーズ以降）に実験条件間で違いはない．

分析対象の 54 対話（Before 条件 33，After 条件 21）で 8 つの主観評定項目それぞれの平均値を算出し，条件間の平均値の差を t 検定により検定した．その結果，Before 条件のシステムの方が「すばやい（有意水準 $p < 5\%$）」と評価され，実験条件の設定がユーザの印象に反映された．しかしそれ以外では After 条件の方が「話しやすい（$p < 2\%$）」とされただけで，相槌タイミングの違いはそれ以外のユーザの主観評定に影響を与えなかった．

しかし分析結果を見ると，Before 条件の評定に関して分散が大きい傾向が見られたので，より詳細に分析するため，被験者をさらに「相槌応答の**重なりに敏感だった**」被験者群（4 人 25 対話）と「相槌応答の**重なりを意識しなかった**」被験者群（6 人 29 対話）とに分けた．分け方は，実験終了後のインタビューにもとづき，実験条件（システム相槌タイミング）の違いに気付き「システム相槌が重なってくることに言及した」被験者群（敏感群）と，「特に言及しなかった」被験者群（意識しない群）とに分けた．

被験者を 2 群に分けた上での印象評定の分析結果を表 2に示した．まず実験条件の実効性を確認するため「すばやさ」項目を見ると，「敏感」被験者群（表 2右欄）では Before 条件の方を「すばやい」と評価したのに対

[4](a) 使いやすい - 使いにくい (b) 話しやすい - 話しにくい (c) 快適だ - 不快だ (d) 好き - 嫌い (e) 親しみが感じられる - 冷たい感じがする (f) 人間的 - 機械的 (g) わかりやすい - わかりにくい (h) すばやい - のろい，の 8 項目．

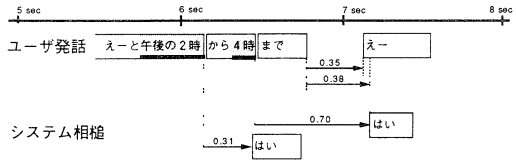

図 4: **ユーザ発話を妨げるシステム相槌応答の例**

して，「意識しない」被験者群（表2左欄）では実験条件間で有意な差が出なかった．インタビュー結果にもとづいて被験者を2群に分けたのは評定値からも適正だったと言える．

それ以外の項目で比較すると，2つの被験者群で正反対の結果が出た．「意識しない」被験者群（表2左欄）は，相槌応答が重なるシステム（Before 条件）での対話を好意的（使いやすい・親しみがある・人間的）に評価する傾向があったのに対して，「敏感」被験者群（表2右欄）では同じ Before 条件での対話を否定的（使いにくい・話しにくい・不快だ・嫌い・機械的）に評価する傾向が見られた．

この結果から，システム相槌の重なりに敏感だった被験者の Before 条件対話には，ユーザに違和感を感じさせる何らかの要因があったのではないか？と考えた．そこで，それらの対話の中から特に評価の低いいくつかの対話の様子を調べた．

すると，ユーザ発話の途中に重ねるシステム相槌がユーザ発話を妨げることがあった．一例を図4に示す．この例では，システムは応答生成ルールに従い，特に遅れることなく相槌を返している．しかし2回目の「はい」がユーザの発声「えー」にぶつかり，ユーザは続けて話せなくなり発声を中断してしまった．対話後のコメントでも「“えー”と言うことで自分が話し続けようとしているのに“はい”に邪魔されて話せなくなる」と述べている．

ユーザが発話している最中にシステムが相槌を打つことは，タイミングよく適切な位置で行われていればユーザの印象を良くするのに対して，タイミングがよくないと（例えばユーザの発声開始直後など）相槌ひとつでユーザの発声を中断させるほどの影響力を持ち，ユーザの印象を悪くすることがわかった．

5 まとめ

人間と音声対話システムによる対話において，システムが情報を受理したことを開示する相槌応答がユーザ発話に重なることで，ユーザにどのような印象を与えるか実験した．分析の結果，システムの相槌応答の重なりを意識しなかった被験者は，発話の途中にも相槌応答を重ねてくるシステムを好意的に評価する傾向があった．しかし反対に，システム相槌の重なりに敏感だった被験者は，同じシステムに否定的な印象を持つ傾向があった．同じふるまいのシステムでも被験者に応じて正反対の印象を与えうること，相槌の重なり方によっては対話の快適性を損なうことがわかった．

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，相川清明リーダ，有益な示唆をいただく対話グループの諸氏，実験の準備と分析にご協力いただいた NTT-AT 社の久保田哲也さん，木間良子さん，実験の被験者に協力していただいたみなさまに感謝いたします．

参考文献

[1] N. M. Fraser and G. N. Gilbert. Simulating speech systems. *Computer Speech and Language*, 5:81–99, 1991.

[2] D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai. A form-based dialogue manager for spoken language applications. In *Proc. ICSLP-96*, 1996.

[3] J. Hirasawa, N. Miyazaki, M. Nakano, and T. Kawabata. Implementation of coordinative nodding behavior on spoken dialogue systems. In *Proc. ICSLP-98*, volume 6, pp. 2347–2350, 1998.

[4] 平沢, 中野, 川端. うなずき・相槌による音声対話システムの理解状態開示. 言語処理学会 第4回年次大会 発表論文集, pp. 182–185, 1998.

[5] 堀口. 日本語教育と会話分析. くろしお出版, 1997.

[6] 岩, 榎本, 大谷京子, 嶋野, 土屋. 日本語地図課題対話における 相手話者発話中の発話開始現象について. 電子情報通信学会技術研究報告 SP98-70, pp. 15–21, 1998.

[7] 川森, 島津. 対話における発話交代の分析. 電子情報通信学会技術研究報告 NLC 95-73, pp. 31–38, 1996.

[8] 西, 北井. 蓄積形音声対話システムにおける発話タイミングと相づちの評価. 電子情報通信学会技術研究報告 SP94-62, pp. 65–70, 1994.

[9] Y. Okato, K. Kato, M. Yamamoto, and S. Itahashi. System-user interaction and response strategy in spoken dialogue system. In *Proc. ICSLP-98*, volume 2, pp. 495–498, 1998.

[10] 小坂. あいづちを中心とした会話音声の呼応関係の分析. 電子情報通信学会技術研究報告 SP87-107, pp. 61–66, 1987.

[11] N. Ward. Using prosodic clues to decide when to produce back-channel utterances. In *Proc. ICSLP-96*, pp. 1728–1731, 1996.

[12] 綿貫, 関, 木山, 荒巻. あいづち位置の考察. 平成10年春季 音響学会講演論文集 3-6-15, pp. 111–112, 1998.