1H1-4

音声対話システムにおける発話の重なり情報を利用した音声認識率低下の予測

Predicting Speech Recognition Performance Degradation Using Utterance Overlapping Information in Spoken Dialogue Systems

中野 幹生[*1] Mikio Nakano　船越 孝太郎[*1] Kotaro Funakoshi　伊藤 敏彦[*2] Toshihiko Itoh　荒木 健治[*2] Kenji Araki　長谷川 雄二[*1] Yuji Hasegawa　辻野 広司[*1] Hiroshi Tsujino

[*1](株) ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan Co., Ltd.

[*2]北海道大学 大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

This paper presents the results of an analysis of user reactions towards system failures in turn-taking in human-computer dialogues. When a system utterance and a user utterance start with a small time difference, the user may stop his/her utterance. In addition, when the user utterance ends soon after the overlap starts, the possibility of the utterance being discontinued is high. Based on this analysis, it is suggested that the degradation in speech recognition performance can be predicted using utterance overlapping information.

1. はじめに

近年多くの音声対話システムが構築されており，その中には実際に運用されているものもある [Zue 00, Komatani 07]．そのようなシステムは，多くの場合，固定的な発話交代メカニズムを採用している．すなわち，ユーザが発話終了後ある閾値以上の長さのポーズを置いた場合にターン（発話の番）をとり，ユーザからのバージイン（割り込み）発話があると即座にターンを譲渡する．音声対話システムのユーザビリティをあげるためには，より柔軟な発話交代を実現する必要がある．

柔軟な発話交代の実現にむけて，これまでいくつかの研究の研究が行われている．ポーズの長さだけではなく，直前のユーザ発話の内容や韻律を用いてターン取得タイミングを判定する研究がある [Sato 02, Ferrer 03, Schlangen 06, Kitaoka 05]．また，バージイン発話（システム発話へのユーザからの割り込み）に対して，ただ単に発話を止めてターンを譲るだけではなく，発話内容も考慮して適切に反応する試みもされている [Ström 00, Rose 03]．

これらの試みにもかかわらず，適切な発話交代はまだ困難である．上記の方法で用いられている素性は，常に正しく求まるわけではない．さらに，人間同士でもいつターンをとるべきかの判断は一致しない [Sato 02]．

したがって，発話交代のタイミングの向上を図るとともに，対話システムが発話交代の失敗に対処できるようにすることが必要である．本稿では，その最初のステップとして，音声対話システムと人間の対話のデータを用い，システムの発話交代誤りに対するユーザの行動の分析を行った．予備的な実験で，発話交代誤りに起因するユーザとシステムの発話の重なりが起きたとき，ユーザが発話を中断することが多いことが判明したが [永野 07]，本研究では，データ量を増やして詳細に分析した．その結果，特定の条件下でユーザの発話が中断されていることが多いことが判明した．

中断した発話は音声認識用文法からはずれる発話が多く，音声認識が誤る可能性が高い．したがって，発話中断の多い状況では，音声認識率が低下することが予想される．我々は，音声認識実験を行うことにより，ユーザの発話が中断されやすい状況下では，音声認識率が低下することを確認した．この結果は，発話の重なり状況が，音声認識率低下の予測，すなわち，音声認識結果の信頼度の推定に役立つことを示唆している．

連絡先: 〒 351-0188 埼玉県和光市本町 8-1 (株) ホンダ・リサーチ・インスティチュート・ジャパン，中野 幹生，E-mail: nakano@jp.honda-ri.com

2. 発話交代失敗に対するユーザの反応の分析

2.1 対話データ

我々は次に述べる 2 つの音声対話システムを用いて収録した対話データ 2 セットを分析した．1 つはレンタカー予約システムで，ユーザが日付，時間，借りる営業所，返す営業所，借りる車のタイプを指定する．もう 1 つはビデオ予約システムで，ユーザが録画したい番組の日付，時間，チャンネル，録画モード（長時間または標準）を指定する．

これらのシステムは 2 つともフレームベースの対話管理を行う．音声認識は，ネットワーク文法による言語モデルを用いることができる Julian[Kawahara 04] とその付属の音響モデルを用いている．レンタカー予約システムの音声認識の語彙サイズは 225 語で，ビデオ予約システムは 198 語である．音声合成には，NTT-IT 社の FinceVoice を用いている．データ収集を行う時には，スタンドマイクとヘッドホンを用いた．対話ごとにマイク入力とシステム出力をステレオファイルに録音した．

対話データセットの内容は以下の通りである．

- セット C: (レンタカー予約)

  23 人（男性 12 人，女性 11 人）の被験者がそれぞれ 8 対話（計 184 対話）行った．各対話では，旅行や出張などの与えられた状況設定に基づき，一回のレンタカー借り出しの予約を行った．対話時間は 3 分とした．134 対話は成功し，38 対話は時間切れになった．12 対話はシステムトラブルにより中断した．

- セット V: (ビデオ予約)

  9 人（男性 13 人女性 4 人）の被験者がそれぞれ 9 対話（計 117 対話）行った．被験者はセット C の被験者とは異

表 1: ユーザ発話とシステム発話の重なりのタイプ別頻度

対話データセット \ 分類	(o1)	(o2)	(o3)	計
C	67	446	7	520
V	46	202	1	249

(o1) ユーザ発話の開始時間がシステム発話の開始時間と終了時間の間である場合

(o2) 一つ以上のシステム発話の開始時間がユーザ発話の開始時間と終了時間の間である場合

(o3) (o1) と (o2) の両方がおこる場合

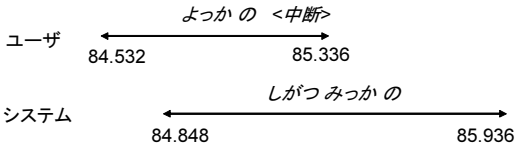

図 1: 発話の重なり時の発話中断（数字は対話開始からの時刻（秒））

なる．各対話では，ユーザは二つの番組のタイマーセットを行った．対話時間は 3 分とした．41 対話では時間内に 2 つの番組の予約に成功し，36 対話では 1 つの番組だけ予約に成功し，34 対話では 1 つも予約できなかった．6 対話では途中でシステムトラブルにより中断した．

両方のシステムとも，バリエーションのある対話を収録するために，対話戦略や発話交代戦略にバリエーションを持たせた．たとえば，確認を行うための音声理解信頼度閾値，発話区間検出のパラメータ，ユーザのバージインに対してシステム発話を即座にとめるかどうかなどの条件を変化した．各被験者の対話の一つ一つにバリエーションを持たせた．本稿では，発話交代失敗の原因よりも現象に焦点をあてるため，これらのバリエーションに関しての詳細は省略する．

対話を収集した後，ユーザ発話とシステム発話の両方を発音通りに書き起こした．発話区間は，アノテーションツールを用いて，300 ミリ秒以上のポーズに基づき，人手で切り出した．タイムスタンプは，ステレオファイルの先頭からの経過時間を用いている．本稿では，各々の発話区間を「発話」と呼ぶことにする．セット C ではユーザ発話数は 3,364，システム発話数は 5,157 であり，セット V ではユーザ発話数は 2,521，システム発話数は 4,522 であった．

2.2 発話の重なり

Raux らが報告しているようにシステムの発話交代失敗にはいくつかの種類がある [Raux 06]．システムがユーザの発話に割り込む場合もあるし，ターンをとるべきなのにとらない場合もある．これらの失敗はさまざまな原因で起こる．たとえば，発話区間検出の失敗や，ユーザがターンを譲ろうとしている意図の理解の失敗などである．

本稿では，ユーザの発話とシステムの発話の重なりを引き起こすような発話交代失敗に焦点をあてる．発話交代失敗の原因については本研究の対象としない．本研究の目的は発話交代を良くすることではなく，発話交代の失敗に対処する方法を見つけることにあるからである．表 1 に発話の重なりのタイプ別頻度をまとめる．

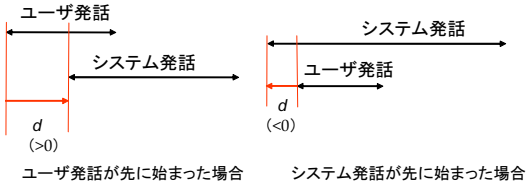

図 2: 開始時刻の差 d

2.3 発話の中断

本稿では，途中で中断された発話を，理由の如何にかかわらず，**自己中断発話**と呼ぶ．我々は，システム発話と重なったユーザ発話が自己中断発話である可能性が高いことを発見した．自己中断発話は，文法的に正しい文ではなかったり，単語のフラグメントを含んだりする場合があるため，音声認識するのが難しい．したがって，自己中断発話を中心に分析を行った．図 1 に，レンタカー予約対話中の自己中断発話の例を示す．

我々は対話の録音のユーザ発話チャンネルだけを聞いて，自己中断発話のアノテーションを行った．セット C の 87 発話が，セット V の 48 発話が自己中断発話であった．これらの中で，61 発話と 38 発話がシステム発話と重なっていた．

自己中断発話の音声認識性能を調べるため，収録に用いた対話システムの音声認識文法と同じものを用いて認識実験を行った．手動で行った音声区間検出を用いているため，収録時の音声認識結果とは異なる．表 2 に示したように，自己中断発話は文法外発話を含むため，単語誤り率が高い[*1]

2.4 発話の中断と発話交代の関係

自己中断発話を発見する方法として，韻律情報を用いることが考えられる [Liu 03]．しかしながら，韻律抽出は完全ではないため，他の方法を検討することが有用である．本研究では，発話交代に関するどのような状況において自己中断発話が起こりやすいかを調査した．

まず，自己中断発話はシステム発話とユーザ発話の開始時刻が近いときに起こりやすいと考えられる．表 3 に，自己中断発話の頻度と，開始時刻の差 d の関係を示す．ここで開始時刻の差 d を以下のように定義する（図 2 参照）．

$$d = st(u) - st(s)$$

ここで，$st(i)$ は発話 i の開始時刻を，u はユーザ発話を，s は u と重なっているシステム発話のうちもっとも時間的に早いものを意味する．この表から，$-0.2s < d < 0.4s$ の場合に，ユーザが自分の発話を中断しやすいことがわかる．d が 0.4s より大きいときは，ユーザはある程度発話を続けているため，その発話を終わらせようとすると考えられる．

次に，自己中断する場合，発話の重なりが始まったらすぐ中断する可能性が高いと考え，システム発話と重なったユーザ発話の終了時刻を調べた．表 4 に発話の重なり開始後どのくらいユーザ発話が継続するか（これを c とする）と自己中断発話の数の関係を示す．c は以下のように定義される（図 3 参照）．

$$c = \begin{cases} et(u) - st(u) & (\text{表 1 の (o1) と (o3)}) \\ et(u) - st(s) & (\text{表 1 の (o2)}) \end{cases}$$

*1 単語誤り率を計算するときの正解データは，書き起こしを単語分割したものを用いているが，文法に現れない単語は 1 モーラごとに単語に分割してしまうため，文法外発話の単語分割には，1 モーラの単語が多く含まれてしまう．このような理由から，文法外発話の単語誤り率が非常に高くなっている．

表 2: 全発話と自己中断発話の音声認識率

セット		すべての発話			自己中断発話		
		文法内発話	文法外発話	全体	文法内発話	文法外発話	全体
C	発話数	2,662	702	3,364	9	78	87
	単語誤り率	22.75%	74.05%	40.23%	12.00%	66.97%	63.13%
V	発話数	1,599	922	2,521	2	46	48
	単語誤り率	13.08%	73.89%	39.69%	0.00%	90.43%	87.39%

表 3: 開始時刻の差 d と自己中断発話の数の関係

d (s)	$-\infty$ – -0.4	-0.4 – -0.2	-0.2 – 0.0	0.0 – 0.2	0.2 – 0.4	0.4 –0.6	0.6 – 1.0	1.0 – ∞
C	2/45	0/7	4/22	15/43	11/56	3/29	4/34	22/284
V	0/17	0/9	10/21	16/57	6/48	3/27	1/12	2/58

(自己中断発話の数)/(重なりのあるユーザ発話の数)

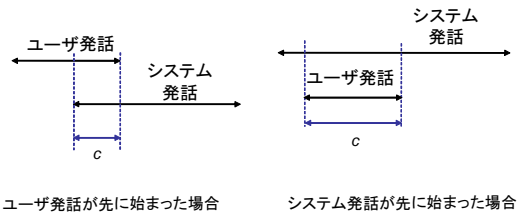

図 3: ユーザ発話の重なり開始後の継続長 c

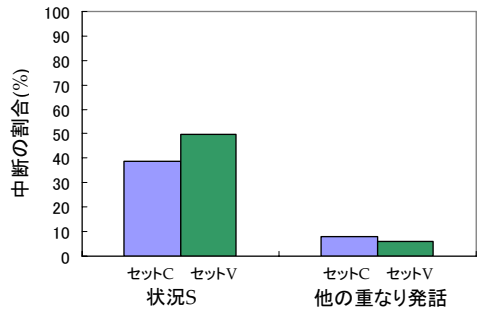

図 4: 状況 S における自己中断の割合

ここで $et(i)$ は発話 i の終了時刻を表す．我々が予測したように，$0.1s < c < 0.6s$ の場合，ユーザ発話は自己中断発話である可能性が高い．

この分析から，自己中断発話は，$-0.2s < d < 0.4s$ かつ $0.1s < c < 0.6s$ の時におこっている可能性が高いと考えられる．この状況を「状況 S」と呼ぶことにする．表 5 および図 4 に d と c の組み合わせと自己中断発話の頻度の関係を示す．

3. 音声認識率の低下の予測

状況 S においては，自己中断発話が起こっている可能性が高いことから，音声認識率が低下することが予想される．表 6 に実際の音声認識率の測定結果を示す．また，図 5 と図 6 に文法外発話の率と音声認識率の比較を図示する．状況 S では

表 5: c と d の組み合わせ毎の自己中断発話の頻度

セット C

d (s) \ c (s)	0.0 – 0.1	0.1–0.6	0.6 – ∞
$-\infty$ – -0.2	0/0	2/12	0/40
-0.2 – 0.4	1/6	**24/62**	5/53
0.4 – ∞	0/44	22/191	7/112

セット V

d (s) \ c (s)	0.0 – 0.1	0.1–0.6	0.6 – ∞
$-\infty$ – -0.2	0/0	0/11	0/15
-0.2 – 0.4	1/2	**26/52**	5/72
0.4 – ∞	0/17	5/47	1/33

各カラム: (自己中断発話の数)/(重なりのあるユーザ発話の数)

文法外発話が多く，音声認識率が実際に低くなっていることがわかる．この結果から，発話の重なり状況が音声認識率の低下の予測に用いられる可能性が示唆された．

4. おわりに

本稿では，音声対話システムと人間との対話における，システムの発話交代失敗に対するユーザの反応の分析結果を提示した．発話交代の失敗により発話が重なった場合に，自己中断が

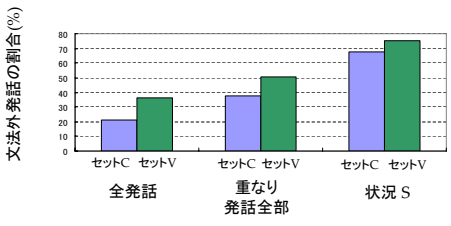

図 5: 文法外発話の割合

表 4: ユーザ発話の重なり開始後の継続長 c と自己中断発話の数の関係

c (s)	0.0 – 0.1	0.1 – 0.2	0.2 – 0.3	0.3 – 0.4	0.4 – 0.5	0.5 – 0.6	0.6 – 0.8	0.8 – 1.0	1.0 – ∞
C	1/50	7/44	10/67	12/66	15/52	4/36	4/75	4/45	4/85
V	1/19	4/19	9/30	13/28	2/17	3/16	2/22	0/17	4/81

(自己中断発話の数)/(重なりのあるユーザ発話の数)

表 6: 状況 S および他の状況での音声認識精度

セット		状況 S			他の重なり発話		
		文法内発話	文法外発話	全体	文法内発話	文法外発話	全体
C	発話数	20	42	62	285	173	458
	単語誤り率	16.67%	107.89%	78.57%	12.72%	66.31%	35.36%
V	発話数	13	39	52	97	100	197
	単語誤り率	9.52%	122.73%	86.15%	8.44%	75.06%	43.14%

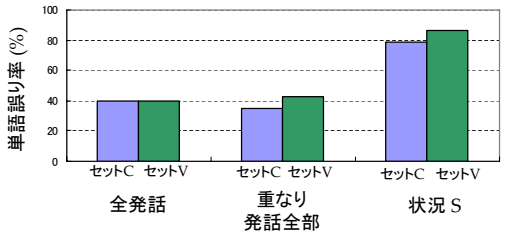

図 6: 単語誤り率

おきやすいことを示し，自己中断の頻度が高い発話の重なり状況を特定した．また，自己中断発話の音声認識が困難であることから，発話の重なり状況を用いて音声認識率の低下を予測できる可能性を示した．これは，システムと人間との対話における誤解の回避につながると期待できる．

今後は，より詳細な分析を行っていく．たとえば，被験者ごとの傾向や対話戦略や発話交代戦略との関係を調べる．また，発話交代失敗の情報を，実際に音声認識の信頼度の向上に用いて行く予定である．本稿では，ネットワーク文法駆動の音声認識を用いたが，統計言語モデルを用いた音声認識でも同様の傾向があるかを調べる．また，本稿では，人手でつけた発話区間情報を用いて分析を行ったが，発話区間検出を自動で行った場合についても研究を進める予定である．

謝辞

システム作成と予備実験に協力していただいた永野由佳氏に感謝します．

参考文献

[Ferrer 03] Ferrer, L., Shriberg, E., and Stolcke, A.: A prosody-based approach to end-of-utterance detection that does not require speech recognition, in *Proc. ICASSP-2003* (2003)

[Kawahara 04] Kawahara, T., Lee, A., Takeda, K., Itou, K., and Shikano, K.: Recent progress of open-source LVCSR engine Julius and Japanese model repository, in *Proc. Interspeech-2004 (ICSLP)*, pp. 3069–3072 (2004)

[Kitaoka 05] Kitaoka, N., Takeuchi, M., Nishimura, R., and Nakagawa, S.: Response Timing Detection Using Prosodic and Linguistic Information for Human-friendly Spoken Dialog Systems, *Journal of The Japanese Society for Artificial Intelligence*, Vol. 20, No. 3 SP-E, pp. 220–228 (2005)

[Komatani 07] Komatani, K., Kawahara, T., and Okuno, H. G.: Analyzing Temporal Transition of Real User's Behaviors in a Spoken Dialogue System, in *Proc. Interspeech-2007*, pp. 142–145 (2007)

[Liu 03] Liu, Y., Shriberg, E., and Stolcke, A.: Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources, in *Proc. Eurospeech-2003*, pp. 957–960 (2003)

[永野 07] 永野 由佳：タスク指向対話における人－ロボット間のロボットジェスチャーと話者交代の分析, 北海道大学工学部情報工学科卒業論文 (2007)

[Raux 06] Raux, A., Langner, B., Bohus, D., Black, A. W., and Eskenazi, M.: Doing Research in a Deployed Spoken Dialog System: One Year of Let's Go! Public Experience, in *Proc. Interspeech-2006 (ICSLP)*, pp. 65–68 (2006)

[Rose 03] Rose, R. and Kim, H. K.: A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems, in *Proc. ASRU-03*, pp. 198–203 (2003)

[Sato 02] Sato, R., Higashinaka, R., Tamoto, M., Nakano, M., and Aikawa, K.: Learning Decision Trees to Determine Turn-Taking by Spoken Dialogue Systems, in *Proc. 7th ICSLP*, pp. 861–864 (2002)

[Schlangen 06] Schlangen, D.: From Reaction To Prediction: Experiments with Computational Models of Turn-Taking, in *Proc. Interspeech-2006 (ICSLP)*, pp. 2010–2013 (2006)

[Ström 00] Ström, N. and Seneff, S.: Intelligent Barge-in in Conversational Systems, in *Proc. 6th ICSLP* (2000)

[Zue 00] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T. J., and Hetherington, L.: JUPITER: A Telephone-Based Conversational Interface for Weather Information, *IEEE Trans. on Speech and Audio Process.*, Vol. 8, No. 1, pp. 85–96 (2000)