2007－SLP－67 (17)
2007／7／21

京都市バス運行情報案内システムにおける実ユーザのふるまいの経時的変化の分析

駒谷 和範　河原 達也　奥乃 博

京都大学大学院 情報学研究科 知能情報学専攻
〒606-8501 京都市 左京区 吉田本町
komatani@kuis.kyoto-u.ac.jp

あらまし
音声対話システムが現実の多様なユーザに対しても適切に動作するためには，ユーザの様々なふるまいへの対処が不可欠である．我々は，京都市バス運行情報案内システムにおける34ヶ月間のデータを用いて，実ユーザのふるまいの経時的変化を分析した．本稿では特にバージインに着目し，発話全体におけるバージインが起こった割合（バージイン率）を分析に加えた．まず，音声認識率やタスク達成率，バージイン率の経時的な変化について調査した．さらに，音声認識率とバージイン率の経時的な変化の間の関係についても調べた．この結果，ユーザがシステムに習熟する過程には，2つの段階を考えることができる．また，各個人のバージイン率により，バージインを伴う発話の音声認識率が異なるという状況も観察された．これにより，バージイン率が音声認識誤りを検出する新たなプロファイルとして利用可能であることも示唆された．

Analyzing Temporal Transition of Real User's Behaviors in Kyoto City Bus Information System

Kazunori Komatani　Tatsuya Kawahara　Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University
Yoshida-Hommachi, Sakyo, Kyoto 606-8501, Japan

Abstract　Managing various behaviors of real users is indispensable for spoken dialogue systems to operate adequately in real environments. We have analyzed various users' behaviors using data collected over 34 months from the Kyoto City Bus Information System. We focused on "barge-in" and added barge-in rates to our analysis. Temporal transitions of users' behaviors, such as automatic speech recognition (ASR) accuracy, task success rates and barge-in rates, were initially investigated. We then examined the relationship between ASR accuracy and barge-in rates. As a result of the analysis, we can model two phases while users get accustomed to the system. We also observed that the ASR accuracy of utterances with barge-ins differed based on the barge-in rates of individual users. The results suggest that the barge-in rate can be used as a novel user profile for detecting ASR errors.

1 はじめに

音声対話システムの性能を向上させるうえで，ユーザのふるまいは考慮されるべき重要な要素である．実ユーザが使用するシステムを設計するには，システムの対話管理のさまざまなユーザへの適応 [1] が必須である．つまり，実際のユーザのふるまいを適切に予測し，それに応じた音声認識や対話管理を行うことで，システムの性能はさらに向上する．具体的には，当該ユーザの音声認識率やそれにより推定される習熟度などに基づき，確認の戦略を変更したり，ヘルプの生成内容 [2] を適応させることが可能となる．我々は今までにも音声対話システムにおけるユーザモデルを考案し，これがシステムの性能を向上させることを示した [1]．しかし，音声対話システムが実用的に使われるためには，長期間にわたる，現実の使用条件下でのユーザのふるまいを知ることが不可欠である．

我々は京都市バスの運行情報案内を音声で行うシステム（075-326-3116）を構築し，運用を続けてきた．本稿では，2002年5月から2005年2月までの34ヶ月間に収集したデータに対して，個々のユーザ（発信者番号）ごとのふるまいを分析した結果を報告する．本稿では特に，音声対話システムに特有の現象であるバージインに着目し，音声認識率やタスク達成率に加えて，ユーザがバージインを行う率も分析に加えた．バージインとは，システムからのプロンプト生成中に音声入力が検出された場合，システムは音声合成を中断し，入力された音声の認識を行うことである．したがって，システムがバージインを許容するように設計した場合，ユーザは冗長なシステムプロンプトを遮ることができ，対話が効率化することが期待されている．これが初心者を含む一般ユーザに対して，実際どのように使われているかを報告する．

ユーザのふるまいの多様性は，個人間の差にとどまらず，同じ個人内でも，慣れによる変化が無視できない．つまり，ユーザはシステムに慣れるにしたがって[1]，どの程度バージインを行うかなどといったふるまいを変えることが予想される．本稿では，ユーザがシステムを使用するにつれて，これらのふるまいがどのように変化したかを調査する．次に，バージインを行う率のユーザによる違いについて調査し，そのうえでこれと音声認識率との関係について報告する．

2 分析対象データ

京都市バス運行情報案内システムにより収集した，2002年5月から2005年2月まで（34ヶ月間）のデータに対して分析を行う．システムは3つのスロット（乗車場所，降車場所，系統番号）を持ち，このうち乗車場所を含む2つの内容が得られると，バスの接近情報を出力する．システムの語彙サイズは，バス停名が652，名所や施設の名前が756である．音声認識はFSAベースで行う．また，ユーザはシステムからのプロンプトの途中に，それを遮って発話することができる（バージイン：barge-in）．もしユーザがシステム発話の内容を既に知っており，それを最後まで聞かなくても次の発話を行える場合には，ユーザはバージインを行うことで，タスクを早く終了させることができる．

システムのログには，コールが行われた時刻や音声認識結果の他に，発信者番号，システムプロンプトが最後まで再生されたか，システムプロンプトの時間などが記録されている．システムプロンプトが最後まで再生されなかった場合，前述のバージインが起きていたとわかる．発信者番号は，ユーザが番号非通知で電話をかけた場合には記録されていないが，全体7,988コールのうち5,927コールで発信者番号が記録されていた．本稿ではこれをもとに，個々のユーザ（発信者番号）ごとのふるまいを分析する．

得られた各コール／各発話に対して，人手でラベルを付与した．ラベルの付与は2名の学生が分担して行った．ラベルの内容は以下である．

1. 発話内容の書き起こし
2. 音声認識結果が誤りかどうか
3. タスクごとの成功／失敗
   タスク成功，タスク失敗，中断，システム調整中
4. その他コメント

2.は，ユーザが発話した内容語が正しく音声認識結果に含まれていた場合，正解とした．3.では，ユーザの音声を人間が聞いたうえで，システムが出力したバスの接近情報がユーザの意図したものであった場合には「タスク成功」とした．

3 ユーザのふるまいの経時的変化の分析

ユーザのふるまいの経時的変化を，以下の3つの尺度において分析する．

- 音声認識率
- タスク達成率
- バージイン率

バージイン率は，当該ユーザの発話数のうち，ユーザがバージインにより入力を行った発話数，として定義している．時間軸として，当該ユーザのある時点までのコール回数を，全コール回数で割った値をx軸とした．したがって$0 < x \leq 1$である．y軸には，

[1] Hofらはユーザがシステムの使用法を忘却する場合も考慮したモデルを提案している [3]．本稿では，スロット数が3という単純なシステムであるため，忘却については考えない．

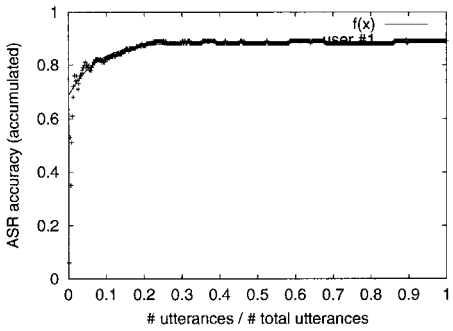

図 1: ユーザ#1 の音声認識率の経時的変化

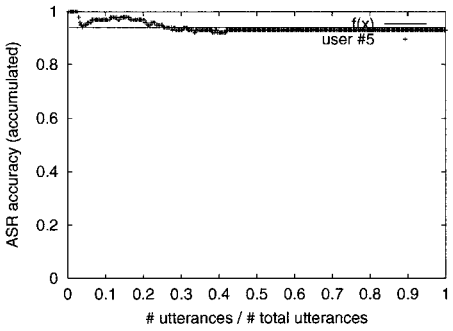

図 2: ユーザ#5 の音声認識率の経時的変化

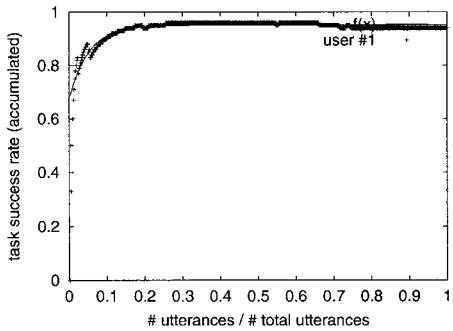

図 3: ユーザ#1 のタスク達成率の経時的変化

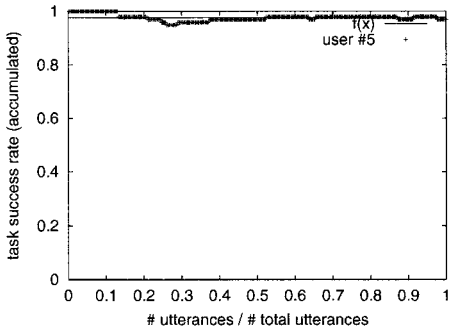

図 4: ユーザ#5 のタスク達成率の経時的変化

そのコールまでの音声認識率，タスク達成率，バージイン率を，それぞれプロットした．

これらの値の変化を以下の関数で近似した．

$$f(x) = c - a \cdot \exp(-bx)$$

c は，ユーザがシステムに十分に慣れた際に，対象とする尺度が収束する値，b は収束の速度，a はこの区間 $(0 < x \leq 1)$ での変化量，におおよそ対応する．これら a, b, c を最小二乗法により算出した．ただし $a \geq 0$ とした．さらに，グラフの概形を表すために，$f(x)$ の変化量がある一定値を下回る際の x を求めた[2]．本稿では，$\frac{df(x)}{dx} = 0.1$ となる x を x_I として求めた．これは x_I 付近で $f(x)$ の変化がほぼ収束することを表す．$f(1)$ は全データに対する当該尺度の平均となる．また $\Delta = f(1) - f(0)$ とし[3]，当該ユーザのこの期間での当該尺度の変化量を表す．$\Delta = 0$ の場合は $f(x)$ の値に変化がないため，x_I は存在しない．

[2] $\frac{df(x)}{dx} = ab \cdot e^{-bx}$ であるため，$ab > 0$ の場合 $f'(x)$ は単調減少する．

[3] 関数近似の結果 $f(0) < 0$ となった場合は，$f(0) = 0$ として算出した．

3.1 音声認識率・タスク達成率の経時的変化

音声認識率，タスク達成率について経時的変化を調べた．ユーザ#1，#5 についてそれぞれ，図 1，図 2 に音声認識率の変化を，図 3，図 4 にタスク達成率の変化を示す．

表 1 に，50 回以上のコールがあった 12 名のユーザに対する経時的変化をまとめた．各関数近似における最小二乗誤差は MSE として示してある．表の左側および中央に，それぞれ音声認識率とタスク達成率の変化を示す値が示されている．音声認識率，タスク達成率の両方において，全区間での音声認識率やタスク達成率の平均 $(f(1))$ は総じて高く，ばらつきも少ない．音声認識率とタスク達成率の変化の相関は大きいこともわかる[4]．ユーザ#1 では徐々に値が増加し，$x = 0.2$ を越えたあたりでほぼ一定値に収束している．一方，ユーザ#5 は両方の場合で初めから値が高く，x が増加しても値に変化はほぼない．このように一部のユーザでは Δ の値が大きく，使用

[4] 同様の傾向は Raux らによっても報告されている [4].

表 1: 多頻度ユーザのふるまいの経時的変化 ($\Delta = f(1) - f(0)$), MSE: 最小二乗誤差

User	音声認識率				タスク達成率				バージイン率			
ID	$f(1)$	Δ	x_I	MSE	$f(1)$	Δ	x_I	MSE	$f(1)$	Δ	x_I	MSE
#1	.88	.20	.25	7.4E-5	.95	.28	.21	1.6E-4	.11	0	-	2.3E-4
#2	.89	.24	.47	2.5E-4	.94	.19	.25	2.8E-4	.19	0	-	1.9E-3
#3	.89	.03	< 0	6.8E-5	.96	.06	< 0	2.1E-4	.60	.60	> 1	6.4E-4
#4	.78	.60	.46	4.5E-4	.89	.89	.52	4.5E-4	.17	0	-	7.2E-4
#5	.94	0	-	3.3E-4	.98	0	-	1.3E-4	.74	.74	.58	4.6E-4
#6	.89	0	-	5.2E-4	.92	.40	.11	7.9E-4	.10	.06	< 0	1.1E-4
#7	.94	0	-	3.2E-4	.93	.09	.08	1.3E-3	.04	.04	.06	1.6E-4
#8	.89	0	-	1.7E-3	.87	.77	.37	1.0E-3	.71	0	-	1.0E-3
#9	.81	.27	.10	4.3E-4	.93	0	-	2.5E-3	.49	.47	.62	4.6E-4
#10	.90	0	-	1.1E-3	1	0	-	1.6E-4	.10	.10	.29	1.3E-4
#11	.72	.20	.17	1.5E-3	.79	.30	.19	2.2E-3	.15	.04	.13	9.8E-4
#12	.79	.37	.21	6.8E-4	.80	0	-	4.7E-3	.23	0	-	2.6E-3

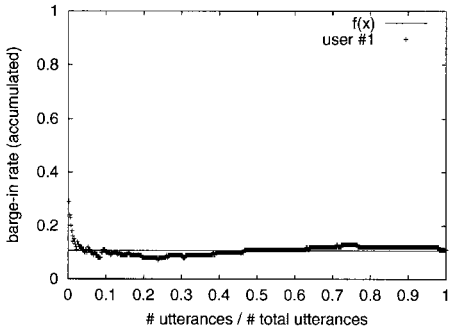

図 5: ユーザ#1 のバージイン率の経時的変化

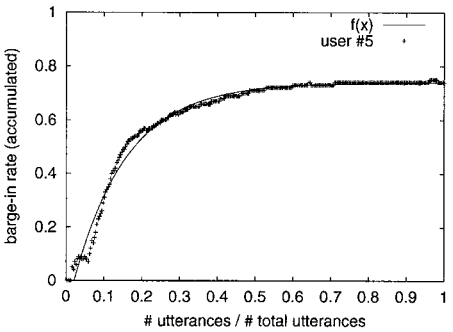

図 6: ユーザ#5 のバージイン率の経時的変化

するにつれて音声認識率やタスク達成率が向上していたことがわかる．つまり，およそ初めからシステムの使い方を知っていたユーザと，システム使用するにつれて徐々にシステムに習熟していったユーザがいることが，定量的に示されている．

3.2 バージイン率の経時的変化

図 5，図 6 にそれぞれ，ユーザ#1，#5 に対するバージイン率の経時的変化を示す．ユーザ#1 の場合は，バージイン率は一定で，ほぼ変化していない．一方ユーザ#5 の場合，システムを使うにつれてバージイン率が上昇している．このように，ユーザによってバージイン率に変化がみられるユーザと，一定のまま変わらないユーザが見られる．

バージイン率においては，表 1 の右側部分からわかるように，音声認識率やタスク達成率とは異なり，$f(1)$ の値のばらつきが多い．これはタスクを遂行する際のユーザのふるまいの多様性を示している．また，#3, #5, #9 など一部のユーザで，バージイン率の大幅な上昇が見られる．一方，残りのユーザでは大きなバージイン率の変化は見られない．このように，ふるまいの変化の度合もユーザによって異なることがわかる．

3.3 バージイン率と音声認識率・タスク達成率との経時的変化の関係

表 1 の結果より，#3, #5 といったバージイン率の変化が大きい（Δが大きい）ユーザについては，音声認識率やタスク達成率におけるΔが小さい．これは，バージイン機能を使いこなせるようになるのは，最初からある程度音声認識率の高かったユーザに限られるという可能性を示唆している．

一方，音声認識率に関するΔが比較的大きいユーザ（#1, #2, #4, #11, #12）については，バージイン率が比較的低い．つまり，システムを使い始め

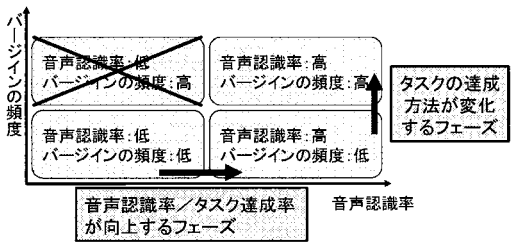

図 7: システムに習熟する過程での 2 つのフェーズ

表 2: バージインの有無による音声認識率

	正解	誤り	合計	正解率
COMPLETE	17,921	3,719	21,640	(82.8%)
BARGE_IN	3,937	4,003	7,940	(49.6%)
Total	21,858	7,722	29,580	(73.9%)

た頃に音声認識誤りが多かったユーザは，システムプロンプトを逐一聞いたうえでタスクを達成するようになる可能性を示唆している．なおユーザ#9 は，音声認識率，バージイン率の双方でΔが比較的大きいが，x_Iの値を見ると，音声認識率の変化が先に収束し，その後にバージイン率の変化が収束していることがわかる．

これらより，ユーザがシステムに習熟する過程には，図7のように，音声認識率／タスク達成率が向上するフェーズと，バージイン率などタスクの達成方法が変化するフェーズがあることが示唆される．ユーザに対する適切なヘルプ [2] を目指すうえで，このフェーズは有用な情報となる．例えば，前者のフェーズではシステム主導の質問や，システムが受理可能な発話パターンを教示するヘルプが有効であるのに対して，後者のフェーズでは「システム応答の間にも割り込んで話すことができます.」のような，タスクの効率的な達成方法に関するヘルプが有効となると考えられる．また，数多くシステムを利用したユーザの中で，図7の「音声認識率：低，バージイン率：高」という状態を経て，システムに習熟していったユーザは見当たらない．つまり，このような状態が検出された場合には，音声認識もしくはバージインの認定に問題がある可能性が考えられる．

4 バージイン率による音声認識誤りの予測の検討

4.1 バージインと音声認識率の関係

次にバージインと音声認識率の関係について調査した．まず，得られた全発話に対して，プロンプトが最後まで再生された場合 (COMPLETE) とバージインがあった場合 (BARGE_IN) における，発話単位の音声認識率を表 2 に示す．全体の発話の 26.8%(7,940/29,580) がバージインにより行われているが，そのうち半数以上が内容語に音声認識誤りを含むものであった．

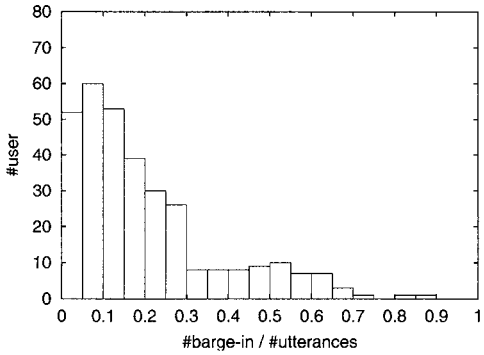

図 8: バージイン率ごとのユーザ数

これには，ユーザが意図して行うバージイン以外に，多数の誤ったバージインが起こっていることを示している．具体的にはまず，携帯電話などによる屋外での利用時に，背景雑音が誤ってバージインと認定され，システムのプロンプトが停止してしまう場合である．また，ユーザの息の音やつぶやきが，誤ってシステムにバージインと認定される場合もある．さらには，ユーザが発話の途中で言い淀むなどすることで，一発話が誤って 2 区間に分割され，その後半部分が更なる誤認識と誤作動を引き起こしている場合もある．上記の背景雑音以外の要因は，ユーザがシステムに適した発話スタイルに習熟していない場合にしばしば起こる．つまり，初心者ユーザは，うまく発話のタイミングが掴めず，システムの誤認識・誤作動を引き起こしている場合が少なからず存在する．

一方，慣れたユーザの中には，バージインを意図して行うことにより，対話を効率的に行っていることが多い．ユーザごとに，全データに対するバージイン率の平均を計算した場合の分布を図 8 に示す．ただしここでは，2 回以上コールを行った 323 ユーザを対象として計数した．図 8 からわかるように，バージインを行う度合はユーザにより大きく異なる．

また表 3 に，一コールごとのバージイン率とその頻度を示す[5]．ここでは全体の一割弱 (528/5,527) のコールで，ユーザは全ての発話でバージインを行ってタスクを達成している．これからも，バージイン

[5] なおここでは，タスク成功に貢献した部分のみを対象に計数しており，例えば何の入力もなく延々とプロンプトが繰り返されたコール（この場合バージイン率は 0 となる）などは含めていない．

表 3: タスク成功部分での発話毎のバージイン率

バージイン率	回数
0.00 - 0.25	3,516
0.25 - 0.50	640
0.50 - 0.75	618
0.75 - 1.00	225
1.0	528
合計	5,527

を有効に使いながら達成されたタスクも相当数あることがわかる．

4.2 バージイン率による音声認識誤りの予測

4.1 節で示されたように，バージインが検出された発話のうち半数は音声認識誤りであった．これは背景雑音やユーザの非習熟によるものが多い．ここで，ユーザごとにバージインを行う度合には差があるため，ユーザごとのバージイン率に基づき，行われたバージインの誤りを検出できる可能性がある．

表 4 に，当該期間全体でのユーザごとのバージイン率と，それに対応する，バージインがあった発話の認識結果との関係を示す．まずバージイン率が 0.8 以上であるユーザの発話は，ほぼ全ての発話でバージインをしていることになるが，これはほとんど雑音などによる誤作動であった．したがって，バージイン率が 0.8 以下のユーザに注目する．バージイン率が高いユーザ，すなわち，日頃からバージインを多く使っているユーザでは，バージインが意図した正しいものである割合が高いと言える．しかし，あまりバージインを行わないユーザ（バージイン率が 0.2 以下）では，起こったバージインが意図したものでない可能性が高く，その発話の認識率は 20%に満たない．

この結果から，例えばバージイン率があるしきい値以下のユーザに対しては，バージインの後に得られた音声認識結果は受理しないといった戦略が考えられる．この戦略により，音声認識誤りの多い部分を正しく棄却することができる．他にも，雑音などによる音声認識誤りを未然に防ぐために，バージイン率が低いユーザに対してはシステムプロンプト中の入力は受け付けないようにするなどの利用法も考えられる．これらより，バージイン率をユーザの特徴を捉えたプロファイルの一つとして用いることができることが，実ユーザのふるまいを分析することにより明らかとなった．

表 4: ユーザごとのバージイン率に対する，バージインがあった発話の認識率

バージイン率	正解	誤り	正解率 (%)
0.0 - 0.2	407	1,750	18.9
0.2 - 0.4	861	933	48.0
0.4 - 0.6	1,602	880	64.5
0.6 - 0.8	1,065	388	73.3
0.8 - 1.0	2	36	5.3
1.0	0	16	0.0
合計	3,937	4,003	49.6

5 おわりに

本稿では，京都市バス運行情報案内システムにより収集したデータにおいて，ユーザのふるまいの経時的変化を分析した．音声対話システムでの対話管理において，ユーザのふるまいのモデル化は非常に重要な要素である．その一部として，バージイン率，音声認識率，タスク達成率の 3 つの尺度について，ユーザ毎の経時的な変化を調査した．今後これらの傾向を特徴として，音声認識誤りの判別や対話管理に活用する方法について検討する．さらに，より多くのデータに対する分析を行い，本稿で述べた傾向の一般性についても検討する予定である．

参考文献

[1] Komatani, K., Ueno, S., Kawahara, T. and Okuno, H. G.: User Modeling in Spoken Dialogue Systems to Generate Flexible Guidance, *User Modeling and User-Adapted Interaction*, Vol. 15, No. 1, pp. 169–183 (2005).

[2] Fukubayashi, Y., Komatani, K., Ogata, T. and Okuno, H. G.: Dynamic Help Generation by Estimating User's Mental Model in Spoken Dialogue Systems, *Proc. Int'l Conf. Spoken Language Processing (INTERSPEECH)* (2006).

[3] Hof, A., Hagen, E. and Huber, A.: Adaptive Help for Speech Dialogue Systems Based on Learning and Forgetting of Speech Commands, *Proc. of 7th SIGdial Workshop on Discourse and Dialogue* (2006).

[4] Raux, A., Bohus, D., Langner, B., Black, A. W. and Eskenazi, M.: Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience, *Proc. Int'l Conf. Spoken Language Processing (INTERSPEECH)* (2006).