音声対話システムにおける ユーザのバージイン率に着目した音声認識誤りの予測

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

komatani@kuis.kyoto-u.ac.jp

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．我々は，京都市バス運行情報案内システム[1]で収集したデータを用いて，34ヶ月間の対話データにおけるユーザのふるまいを分析してきた [1, 2]．これにより，タスクを達成する際のターン数に個人差があることや，バージイン (barge-in) により入力された発話が音声認識誤りかどうかを予測するのに，ユーザがバージインを行う度合が有効である可能性を確認した [1]．さらに，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ことを実験的に示した [2]．このように，現実の使用条件下でのユーザのふるまいは多様である．これを適切に各ユーザのプロファイルとしてモデル化したうえで，音声認識や対話管理を適応させることは，システムの性能向上につながる [3]．

本稿では，新たなユーザプロファイルとしてユーザがバージインを行う率（バージイン率）に注目する．バージインとは，システムからの応答生成中にユーザが発話を行う現象である．この場合，システムは音声合成を中断し，入力されたユーザ音声の認識を行う．バージインは，ユーザがそのシステムにどの程度慣れているか [3] や，ユーザがシステムをどの程度擬人化して扱っているかなど，対話管理に有用な特徴を有する．さらに，バージインの有無は，ユーザの発話のタイミング情報から得られるため，ほぼ誤りなく取得でき，オンラインでバージイン率を推定することも容易である．我々は以前，各ユーザの平均バージイン率と音声認識の成否に相関がある傾向を示した [1]．

本稿では，これをさらに発展させ，ユーザのふるまいの経時的変化を考慮してバージイン率を計算する．

[1](075)326-3116 での運用は 2007 年 3 月末で終了した．現在新システムを IP 電話ベースで試験運用中である (050-5539-9669)．

表 1: バージインの有無による音声認識率 [1]

	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%)

これにより，バージイン率の計算のオンライン化を可能としたうえで，音声認識の成否の予測精度を向上させることを狙う．

2　これまでの知見

2.1　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 [1, 2]．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，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／各発話に対しては，発話内容の書き起こしや，音声認識結果が誤りかどうかなどのラベルを人手で付与した．

2.2　平均バージイン率による音声認識誤りの予測

得られた全発話に対する，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

表 2: 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音声認識率 [1]

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

場合 (BARGE_IN) の，発話単位の音声認識率を表 1 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．これは背景雑音やユーザのシステムへ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 2 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率の関係を示す．

2.3 ユーザのふるまいの経時的変化

我々は，ユーザのふるまいの経時的変化を，音声認識率，タスク達成率，バージイン率の 3 尺度において分析した [2]．ここではバージイン率に絞って述べる．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定義した．

図 1 に，あるユーザの，時間変化に伴うバージイン率の変化を示す．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，そのコールまでのバージイン率を，それぞれプロットした．このように，ユーザのふるまいはシステムに慣れるにつれて変化するため，ユーザ間の違いだけでなく，一ユーザの中でもふるまいの経時的変化をモデル化する必要性が示されている．

3 オンライン検出を指向したバージイン率の計算

バージイン率をオンラインでユーザプロファイルとして利用する場合を考える．つまり，表 2 のように，

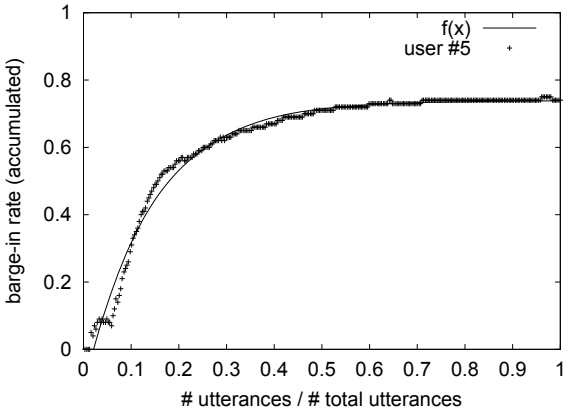

図 1: あるユーザのバージイン率の経時的変化 [2]

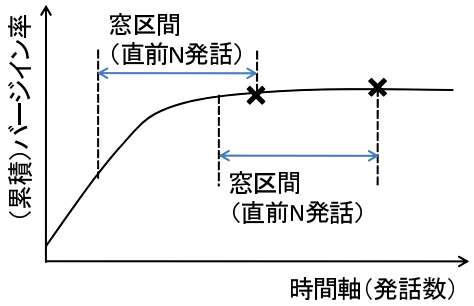

図 2: ある発話時点での窓区間内の平均や分散の算出

データ全体でバージイン率を計算するのではなく，その時点までの直前 N 発話を対象として平均を計算し，プロファイルとして利用する．以降，この直前 N 発話を窓区間と呼ぶ．計算の模式図を図 2 に示す．これにより，対話の各時点におけるバージイン率を定義し，図 1 のように経時的に変化するユーザのふるまいに対応する．

さらに，この窓区間内での，バージイン率の分散（標準偏差）も合わせて考える．これは，この分散が小さい場合，つまりバージイン率が収束してあまり変化しない場合は，安定したユーザプロファイルと捉えられることを意図する．一方，分散が大きい場合は，ユーザのふるまいが一定せず，不安定であることを示す指標といえる．

3.1 実験的検証

バージインがあった発話の音声認識の成否を，各時点での当該ユーザのバージイン率などユーザプロファイルを用いて予測する．予測にはロジスティック回帰

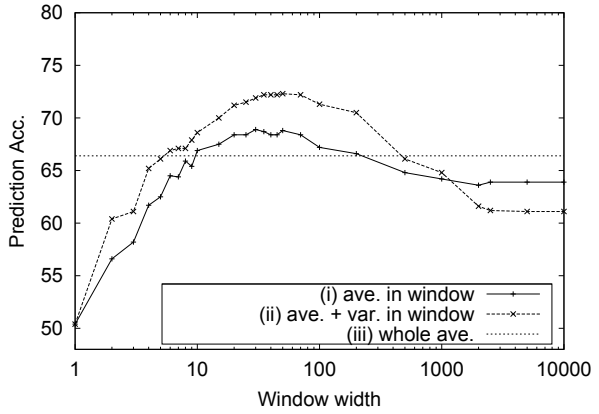

図 3: 全バージイン発話に対して窓幅を変えた際の音声認識誤り予測精度の変化

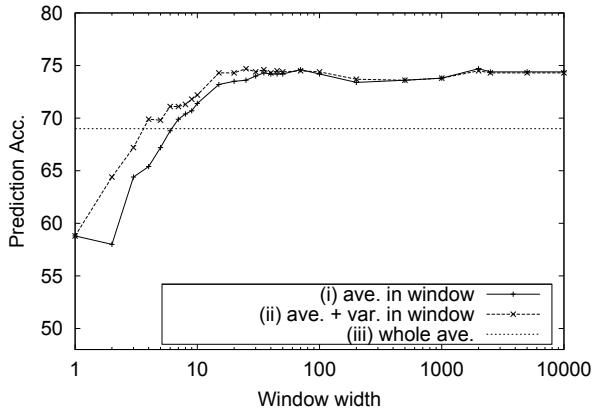

図 4: 最小コール数 10 回以上のユーザに対して窓幅を変えた際の音声認識誤り予測精度の変化

を用いる．つまり，音声認識が正しい確率を P_{ASR} とすると，

$$P_{ASR} = \frac{1}{1+\exp(-(a_1x_1 + a_2x_2 + b))}$$

となるように，学習データに対して最適なパラメータ a_1, a_2, b を求める．ここでは入力 x_1, x_2 として，窓区間内のバージイン率の平均と標準偏差を考えた．パラメータ推定は 10-fold cross validation で行った．

まず，バージインがあった全ての発話（7940 発話）を対象として，窓幅を変えてバージイン率を計算した場合の，音声認識誤り予測精度の変化を図 3 に示す．図中の 3 つの条件は，それぞれ下記のように入力を変えた場合の，音声認識誤り予測精度の平均である．

(i) ave. in window: 各窓幅内の発話のバージイン率の平均のみを入力とする．

表 3: 音声認識誤り予測精度の最大値とその時の窓幅

	(i)	(ii)	(iii)	Maj.
全発話	68.9% (w=30)	72.3% (w=50)	66.4% (-)	50.4% (-)
最小 10 コール	74.6% (w=70)	74.7% (w=25)	69.1% (-)	58.8% (-)

() 内は予測精度が最大となった時の窓幅 (w)

(ii) ave. + var. in window: 窓幅内の発話のバージイン率の平均と標準偏差をともに入力とする．

(iii) whole ave.: 全対話を通じてのバージイン率の事後的な平均を，各発話時点での入力とする．表 2 のデータに相当する．

さらに，対象とする発話を，コール数が 10 回以上であったユーザによる発話に限定した場合（6216 発話）の予測精度の変化を，図 4 に示す．ユーザの最小コール数で限定したのは，利用できる履歴（発話数）が少ないユーザを取り除いた場合の精度を検証するためである．1 コール内の発話数はおおよそ 2 から 5 程度である．

また，図 3，図 4 における，各条件での予測精度の最大値とその際の窓幅を表 3 に示す．その際の推定パラメータ値は以下のとおりである．

全発話, 条件 (i) (窓幅 30)
$a_1 = 3.08, b = -1.60$

全発話, 条件 (ii) (窓幅 50)
$a_1 = 3.05, a_2 = -7.54, b = -1.04$

最小 10 コール, 条件 (i) (窓幅 70)
$a_1 = 4.16, b = -1.65$

最小 10 コール, 条件 (ii) (窓幅 25)
$a_1 = 4.13, a_2 = -3.66, b = -1.50$

なお，表 3 中の条件 “Maj.” は Majority baseline を表し，全ての発話を正解（または誤り）に分類した場合の精度である．

3.2 考察

経時的変化のモデルの必要性：図 3 図 4 の両方で，窓幅を適切な範囲（数十発話）に設定すると，全体の平均バージイン率を用いる場合（条件 (iii)）よりも，インクリメンタルにバージイン率を計算した場合（条件 (i)）の方が予測精度がよい．これは，図 1 に示したように，推定すべきバージイン率が全体で一定ではな

く，経時的に変化するためである．したがって，バージイン率の経時的変化を考慮したモデルの必要性を示している．

必要な窓幅：3.1 節の結果より，窓幅を広げていった場合には，およそ 30 発話ほどで音声認識率の予測精度の上昇が飽和している．つまり，バージイン率をユーザプロファイルとして用いる際には，本実験条件では少なくとも直前 30 発話ほどのバージイン率の平均をとればよいことが示されている．

標準偏差を併用する効果：図 3 と図 4 の結果では，前者の方が窓幅に満たない回数しか発話していないユーザが多い．この場合は，履歴の数がユーザプロファイルとして用いるには十分ではない．また，図 4 では，窓幅が狭い場合には標準偏差を併用した場合（条件 (ii)）の方が精度が高いが，窓幅を十分に大きくすると標準偏差の有無は精度にほぼ影響しない．この結果より，入力とするバージイン率にノイズが大きいと思われる場合には，標準偏差を併用する効果が見られた．

4　まとめと今後の課題

本稿では，各ユーザのバージイン率をプロファイルとして用いて，音声認識の成否の予測を試みた．バージインの有無は，ほぼ誤りなく取得できる特徴であるため，これを用いて音声認識誤りの予測精度が向上すれば有用である．本稿では，各ユーザのプロファイルとして，各時点の直前 N 発話の窓をかけて，バージイン率の平均やその標準偏差を計算して用いた．数十発話ほどの窓幅を設定してバージイン率の平均を逐次計算することで，全体のバージイン率の事後的平均を用いる場合と比べ，予測精度が向上した．これにより，経時的変化のモデルの必要性が定量的に示された．

本稿で述べた音声認識の成否の予測は，経時的変化を考慮したバージイン率と，音声認識率との間に相関があることを利用している．しかし，ユーザのふるまいの経時的変化の分析の結果,「バージインの頻度は低いが音声認識率は高い」という段階を経て，ユーザのふるまいは変化していくという知見を我々は得ている（図 5）[2]．したがって，ある時点での当該ユーザの音声認識率をオンラインで推定できれば，本稿で扱ったモデルをさらに高精度化できる．今後，システムの明示的確認に対するユーザの応答を利用した，音声認識率のオンライン事後推定 [4, 5] を行うなどして，本モデルの高精度化を進める．またこの結果を用いて，

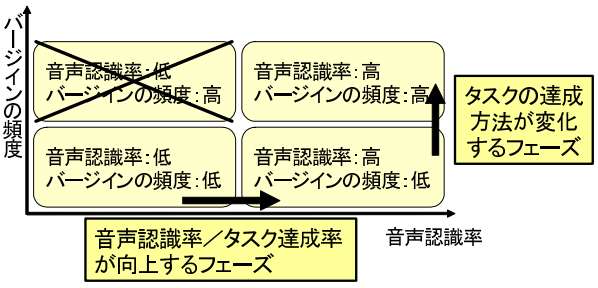

図 5: システムに習熟する過程での 2 つのフェーズ [2]

図 5 で示されているフェーズに応じたヘルプ生成にもつなげる．

参考文献

[1] 駒谷和範, 河原達也, 奥乃博: 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分析, 言語処理学会第 12 回年次大会発表論文集, pp. 42–45 (2006).

[2] 駒谷和範, 河原達也, 奥乃博: 音声対話システムにおけるユーザのふるまいの経時的変化の 分析, 言語処理学会第 13 回年次大会発表論文集, pp. 147–150 (2007).

[3] 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).

[4] Sudoh, K. and Nanano, M.: Post-Dialogue Confidence Scoring for Unsupervised Statistical Language Model Training, *Speech Communication*, Vol. 45, pp. 387–400 (2005).

[5] Bohus, D. and Rudnicky, A.: Implicitly-supervised Learning in Spoken Language Interfaces: an Application to the Confidence Annotation Problem, *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, pp. 256–264 (2007).