音声対話システムにおける 各ユーザの利用履歴を活用したバージイン発話のエラー検出

駒谷 和範　　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

音声対話システムにおいて，音声認識誤りは最大の問題である．この誤りに起因するシステムの誤動作を防ぐために，音声認識の信頼度など発話レベルの特徴 [1] に加えて，対話レベルの特徴を用いた研究も行われている [2, 3]．とりわけ実ユーザに一般公開されたシステムでは，初心者を含む多様なユーザによる多様な発話による誤りを正しく検出する能力が必要である．また公開されたシステムではユーザがシステムを繰り返し使用する場合があり [4]，この場合は，各ユーザごとにモデルを作ることで誤り検出性能が向上する [5]．つまり繰り返し利用される音声対話システムにおいて，ユーザ情報は非常に有用な情報である．

一方，対話システムという状況を生かして，信頼できる発話の認識結果を教師信号として学習に生かす試みが行われている．例えば須藤らは，対話終了時に確定するコンセプトは正しいと仮定し，それを入力した発話の音声認識結果も正しいと仮定した [6]．Bohus らはシステムによる明示的確認に対する肯定／否定を手掛かりとして，それらの発話を信頼度付与の際の教師信号として用いている [7]．このように，ある音声認識結果を対話の後に正しいとみなせるなら，それらを教師信号とした機械学習を導入できる．このアプローチでは，音声対話システム開発で最もコストがかかる発話への正解ラベル付与の労力を必要としない．我々は後者の Bohus らの研究と同様に，システムによる明示的な確認に対する肯定／否定応答結果を手がかりとして，事後的にそのユーザの音声認識率を推定し，発話の取捨選択に用いている [8]．

本稿では，この推定された音声認識率や，バージイン率，つまりユーザがそれまでにどれくらいシステム発話に割り込んだか，に加え，さらに従来からよく用いられる音声認識信頼度を用いて，音声認識結果の誤りの予測を行う．本研究の特徴は以下の 2 点である．

1. ユーザの利用履歴を当該ユーザのプロファイルとして用いる．具体的には当該ユーザのバージイン率や音声認識率を特徴として予測に活用する．
2. オンラインでの検出を指向する．つまり事後的にではなく，システム公開中に人手による介入なしに，上記のプロファイルの獲得を目指す．

以前の報告で我々は，推定音声認識率の計算法を定義し，これをバージイン率とともに用いて予測精度が向上することを示した [8]．本稿ではこれらにさらに音声認識信頼度を加えて，本アプローチの有効性を実験的に検証する．これにより，ユーザごとのバージイン率や音声認識率が，ユーザのプロファイル（事前情報）として有用かどうかを検証する．

2　バージイン発話の誤りとその予測

本稿ではバージイン発話の誤りの予測をタスクとする．この予測精度を，人手での書き起こしを必要とせずに向上させるのが最終目的である．予測精度の向上により，従来の信頼度を用いた取捨選択のように，認識誤りを含む音声認識結果による誤動作や，冗長な確認を防げる．なおバージイン発話とは，ユーザがシステムの発話に割り込んで行った発話である．

我々の収集したデータにおいては，主に音声認識誤りに起因する，バージイン発話の解釈誤りが多く見られた．表 1 に，プロンプトが最後まで再生された場合 (”バージイン無”) とバージインがあった場合 (”バージイン有”) の，発話単位の音声認識率を示す．ここでは一発話中の内容語の認識結果が全て正しい場合のみを正解としており，一部でも内容語に認識誤りが含まれる場合は誤りとして計数している．表 1 より，全体の発話の 26.2% (7,193/27,499) 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んでいたことがわかる．

この傾向は，ユーザがシステム発話に割り込む際には，システムプロンプトの終了を待ってから話し始めるよりも言い淀みが起こりやすいという Rose らの調査結果とも合致する [9]．ユーザの言い淀みやそれによる発話断片は，音響的にはほぼユーザの発話そのものであるため，ユーザ発話と雑音を識別する GMM[10] など，音響レベルの特徴のみを用いて棄却するのは難しい．これらのエラーの検出には，音響信号や音声認識結果など，一発話からボトムアップに得られる情報以外の情報が必要である．

この問題に対して，我々はユーザの利用履歴から得

表 1: バージインの有無による音声認識率

音声認識結果	正解	誤り	合計	(精度)
バージイン無	16,694	3,612	20,306	(82.2%)
バージイン有	3,281	3,912	7,193	(45.6%)
合計	19,975	7,524	27,499	(72.6%)

られる情報を，各ユーザのプロファイルとして活用する．具体的には，各ユーザのある時点までのバージイン率と音声認識率を用いる．バージイン率は，直感的には，ユーザがそのシステム，特にそのバージイン機能への習熟度合に対応する．つまり，バージインを頻繁に使ってタスクを多数遂行しているユーザは，バージイン発話のエラー率が低いという傾向 [4] を利用している．また，各ユーザの音声認識率もシステムへの慣れを表す指標となる．つまり，習熟したユーザの音声認識率は高いという傾向に対応する [4, 11]．一方で，バージイン率と音声認識率は必ずしも同時には向上しないことも確かめられている [4]．実際，熟練して音声認識率が高くなってもバージイン率が低いユーザも存在し，全ての熟練ユーザがバージインを頻繁に行うとは限らない [4]．このため，バージイン率とともに，当該ユーザの音声認識率を用いて，多様なユーザの熟練の度合を表現し，誤り予測のための事前情報としての有効性を検証する．

3 各ユーザの利用履歴を用いたエラー予測

本研究では以下の 3 種類の情報を用いて，ユーザ発話のエラーを予測する．

1. それまでの当該ユーザのバージイン率
2. それまでの当該ユーザの音声認識率
3. その発話の音声認識信頼度

各ユーザの利用履歴としてバージイン率と音声認識率を用いる．具体的には，コーパス中の各時点においてロジスティック回帰により発話の正解不正解を予測する．ここでは正解に 1，誤りに 0 を割り当てている．

$$P = \frac{1}{1 + \exp(-(a_1 x_1 + a_2 x_2 + a_3 x_3 + b))} \quad (1)$$

ここでの x_iとして，そのユーザのそれまでのバージイン率と音声認識率，およびその発話の音声認識信頼度を用いる．パラメータ a_i, b は評価データに対する 10-fold cross validation により推定する．

バージイン率は，現発話を含めて，当該ユーザのそれまでの発話のうち，バージインが行われた発話の割合とする．ここで，バージイン率はユーザのシステムへの慣れに従って経時的に変化する [4] ため，一定の窓幅内でのバージイン率を計算する．つまり窓幅を N とすると，直近 N 発話のみを使ってバージイン率を計算し，変化前の古い履歴を無視する [5]．窓幅がそのユーザの全発話数より大きい場合は，そのユーザのそれまでの全発話を用いる．このため，窓幅が一ユーザによる最大発話数 2838 発話を越えた場合は，それ以前の発話の全てを使った場合と等価になる．

音声認識率は発話単位で計算する．つまり発話中のタスク遂行に必要な内容語が全て正しく認識されている場合に正解とし，誤りのある内容語を含む発話は不正解とする．また音声認識率は，当該発話の一発話前までの，そのユーザの全発話を用いて計算する．各ユーザが最初にシステムを利用した場合の音声認識率は，一発話前がないため 0 とした．さらに音声認識率の計算ではバージイン率の場合のような窓は設定しなかった．これは，音声認識率に窓を設定しても精度は特に向上しなかったためである [8]．この理由は，ユーザの音声認識率はバージイン率よりも早く収束するため [4]，バージイン率と比較すると音声認識率には大きな変化がなかったためと考えられる．

ここでは音声認識率として以下の 2 種類を用いる．

1. 正解音声認識率
2. 推定音声認識率 [8]

正解音声認識率は音声認識結果の書き起こしから計算したものであり，音声認識率を使う場合の性能の上限を調べるために用いる．つまりオンラインでは得られない．一方，推定音声認識率は，実行時のユーザの肯定／否定応答を利用して，書き起こし無しでユーザのそれまでの音声認識率を推定する [8]．正解音声認識率との相関係数は 0.806 と比較的高い値が得られた．この音声認識率の推定値は肯定／否定応答が行われるたびに更新する．肯定／否定応答以外の発話での推定音声認識率は，直近の肯定／否定応答で計算した推定音声認識率とみなす．

音声認識信頼度は発話単位で計算されるものを用いる．本稿では Nuance 社[1]により開発された Voice Web Server (VWS) 付属の音声認識器が出力した信頼度を 100 で割ったものを用いる．表 2 にバージイン発話に対する信頼度の分布を示す．単純にしきい値を用いて $CM > 0.516$ で受理とすると，この信頼度だけでも 90.8%と高精度な判別が可能である．

4 実験的検証

4.1 対象データ

評価用データとして，京都市バス運行情報案内システム [12] で収集したデータを用いた．これは，ユーザの要求するバスがあとどれくらいで到着するかを音声で出力するシステムである．本システムは，電話を

[1] http://www.nuance.com/

表 2: バージイン発話 (7,193 発話) に対する信頼度の分布

信頼度	正解	誤り	(%)
0.0 - 0.1	0	1491	0.0
0.1 - 0.2	0	69	0.0
0.2 - 0.3	0	265	0.0
0.3 - 0.4	0	708	0.0
0.4 - 0.5	241	958	20.1
0.5 - 0.6	639	333	65.7
0.6 - 0.7	1038	68	93.9
0.7 - 0.8	1079	20	98.2
0.8 - 0.9	284	0	100.0
0.9 - 1.0	0	0	–
合計	3281	3912	45.6

通じて一般市民からアクセス可能であった．システムは誤った情報提供を行わないように，全てのユーザ発話に対して常に明示的確認をするという最も保守的な戦略を採っていた．

実験には全 7,988 コールのうち，電話番号が記録されていない 2,061 コールとシステムのデバッグのための 933 コールを除き，671 名のユーザから得た 4,919 コール，合計 27,499 発話を用いた．このうち，7,193 発話がバージイン発話であった（表 1）．評価のため各発話は人手で書き起こされており，その内容語が正しいかどうかも人手で正解ラベルを与えた．

各コールの電話番号は基本的に記録されているため，それぞれの電話番号を一ユーザとみなした．タスクの性質上，大多数の電話番号が携帯電話のものであり，一般に携帯電話を他人と共有することは少ないため，この仮定は概ね妥当と考える．

4.2 実験条件

バージイン発話 7,193 発話に対するエラー予測の精度を調べた．表 3 に，式 1 に入力する特徴 x_i の集合を変えた場合の予測精度を降順で示す．バージイン率を用いる条件では，精度が最も高くなった窓幅 w の値と，その時の予測精度を載せている．また式 1 のロジスティック回帰式の出力値の分布の良さを調べるために MAE も計算した．MAE は Mean Average Error の略であり，予測値と正解との 1 発話あたりの誤差の平均である．具体的には以下の式で計算する．

$$MAE = \frac{1}{m}\sum_{j}^{m} |\hat{X}_j - X_j| \qquad (2)$$

ここで m は全発話数，X_jは j 番目の発話への出力値，$\hat{X}_j$は人手で与えた正解ラベル (0 または 1) である．(12) 単純手法 では，予測精度として Majority baseline，つまり全て 0（不正解）を出力した場合の精度を計算し，MAE には平均認識率（つまり 0.456）を常に出力値とした場合を記載している．

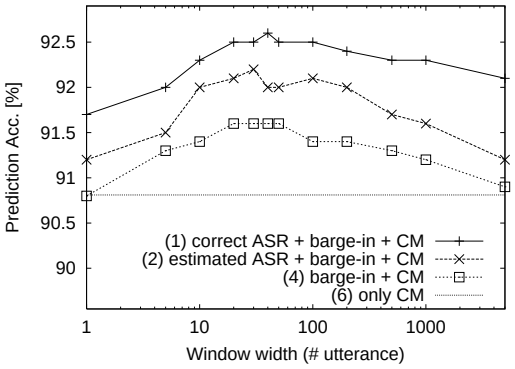

図 1: バージイン率計算の窓幅を変化させた際の予測精度

4.3 実験結果

表 3 の結果を以下で順に説明する．まず用いている音声認識信頼度（CM）自体が表 2 で示されるように高精度であるため，これを用いた条件 (1)〜(6) の予測精度が高い．MAE の値も小さく，出力値の性能の高さも示されている．

次に，条件 (6) と条件 (1)〜(5) を比較すると，音声認識信頼度に加えてバージイン率や音声認識率を利用することで，精度が向上している．つまり，これらのユーザの利用履歴は一発話から得られる音声認識信頼度とは異なる情報源であり，音声認識信頼度とともに用いた場合でも精度向上に貢献することが示されている．

図 1 にさらに詳しい状況を示す．ここでは条件 (6) と条件 (1)(2)(4) をプロットしている．音声認識信頼度（CM）のみを用いる条件 (6) では，バージイン率を用いていないため，窓幅が変化しても予測精度は一定である．条件 (1)(2)(4) はバージイン率の窓幅によって予測精度が変化し，おおよそ窓幅 w が 30〜40 で予測精度が最大になっている．これは以前の報告 [5, 8] とほぼ一致し，1 対話の平均発話数が 5 前後だとすると，同一ユーザがおよそ 10 回弱システムを利用すれば，その利用履歴がプロファイルとして有効に働くことを示している．

条件 (2) と (4) を比較すると，条件 (4) の CM+バージイン率に，推定音声認識率を加えることで精度が向上している．これにより推定音声認識率も精度向上に貢献していることがわかる．また条件 (1) と (2) を比較すると，正解音声認識率を使用した条件 (1) の方が，推定音声認識率を用いた条件 (2) よりも予測精度が高い．つまり音声認識率の推定方法を改善すれば，条件 (1) を上限として，推定音声認識率を使った場合の精度の向上の余地があることがわかる．

表 3: 各条件での最高予測精度とその時の窓幅 w

条件 (入力)	窓幅 w	予測精度 (%)	MAE
(1) CM+バージイン率+正解音声認識率	w=40	92.6	0.112
(2) CM+バージイン率+推定音声認識率	w=30	92.2	0.119
(3) CM+正解音声認識率	-	91.7	0.121
(4) CM+バージイン率	w=30	91.6	0.126
(5) CM+推定音声認識率	-	91.2	0.128
(6) CM	-	90.8	0.134
(7) バージイン率+正解音声認識率	w=50	80.0	0.312
(8) バージイン率+推定音声認識率	w=50	77.7	0.338
(9) 正解音声認識率	-	72.8	0.402
(10) バージイン率	w=30	71.8	0.404
(11) 推定音声認識率	-	57.6	0.431
(12) 単純手法	-	54.4	0.496

MAE: Mean Absolute Error

5 まとめ

本稿では，ユーザごとの利用履歴から得られる情報をプロファイルとして用いて，バージイン発話のエラーを予測する手法について述べた．推定音声認識率を使用する場合は，バージイン率，音声認識信頼度ともにオンラインで得られるため，発話の書き起こしなど，人手によるラベル付けは不要である．本稿ではこのような当該ユーザの利用履歴から得られる情報が，音声認識信頼度とともに用いる場合でも依然有用であり，発話のエラー予測に有効な手がかりであることを実験を通じて示した．

本手法は，図 1 の結果からもわかるように，おおよそ 10 回以上繰り返し使用されるシステムにおいて有効である．またユーザ ID が所与であることも仮定している．さらに音声認識率の推定部分 [8] では，各発話に対して必ず明示的確認を行うという対話戦略を採るシステムであることを仮定し，また肯定／否定応答を音声認識結果のまま単純に全て正解とするなど，比較的単純な手法が採られている．今後の課題として，これ以外の対話戦略を採るシステムでも有効となる，より高精度な推定音声認識率の計算方法の開発が挙げられる．

謝辞 評価データは京都市バス運行情報案内システムにより収集されたものである．データ収集にご尽力くださった京都大学教授の河原達也先生に感謝します．本研究の一部は科研費の支援を受けた．

参考文献

[1] Komatani, K. and Kawahara, T.: Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output, *Proc. COLING*, pp. 467–473 (2000).

[2] Litman, D. J., Walker, M. A. and Kearns, M. S.: Automatic Detection of Poor Speech Recognition at the Dialogue Level, *Proc. ACL*, pp. 309–316 (1999).

[3] Walker, M., Langkilde, I., Wright, J., Gorin, A. and Litman, D.: Learning to Predict Problematic Situations in a Spoken Dialogue System: Experiments with How May I Help You?, *Proc. NAACL*, pp. 210–217 (2000).

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

[5] Komatani, K., Kawahara, T. and Okuno, H. G.: Predicting ASR Errors by Exploiting Barge-In Rate of Individual Users for Spoken Dialogue Systems, *Proc. INTERSPEECH*, pp. 183–186 (2008).

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

[7] 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).

[8] 駒谷和範, Rudnicky, A. I.: 音声対話システムにおける暗黙的な教師信号に基づく音声認識率の推定とそれを用いたエラー予測, 情報処理学会研究報告, SLP-78-3 (2009).

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

[10] Lee, A., Nakamura, K., Nisimura, R., Saruwatari, H. and Shikano, K.: Noice Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs, *Proc. ICSLP*, pp. 173–176 (2004).

[11] Levow, G.-A.: Learning to Speak to a Spoken Language System: Vocabulary Convergence in Novice Users, *Proc. 4th SIGdial Workshop on Discourse and Dialogue*, pp. 149–153 (2003).

[12] 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).