Vol.2009-SLP-78 No.3
2009/10/23

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

駒谷 和範[†1]　Alexander I. Rudnicky[†2]

個々のユーザのふるまいのモデル化は，一般市民に向け公開され繰り返し使われる音声対話システムの性能を向上させるのに有望な方法のひとつである．我々は，システムの明示的な確認の後に続くユーザ応答を用いて，ユーザごとに,「暗黙的な教師信号に基づく」推定音声認識率を計算する．この推定音声認識率を，そのユーザのシステムへの慣れを表すバージイン率と統合し，バージイン発話の誤り予測を行う．評価実験により，本稿で定義した推定音声認識率が，この誤り予測精度の向上に有用であることを示す．なおここで用いた推定音声認識率とバージイン率はともに実行時に計算可能であるため，人手による正解付与作業なしに誤り予測性能を向上させるのに用いることができる．

Predicting Barge-in Utterance Errors by using Implicitly Supervised ASR Accuracy and Barge-in Rate per User

Kazunori Komatani[†1] and Alexander I. Rudnicky [†2]

Modeling of individual users is a promising way of improving the performance of spoken dialogue systems deployed for the general public and utilized repeatedly. We define “implicitly-supervised” ASR accuracy per user on the basis of responses following the system’s explicit confirmations. We combine the estimated ASR accuracy with the user’s barge-in rate, which represents how well the user is accustomed to using the system, to predict interpretation errors in barge-in utterances. Experimental results showed that the estimated ASR accuracy improved prediction performance. Since this ASR accuracy and the barge-in rate are obtainable at runtime, they improve prediction performance without the need for manual labeling.

1. はじめに

音声対話システムにおいて，音声認識結果は最大の入力情報であり，したがってその誤りは最大の問題である．その誤りの受理によ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[2] に加えて，対話レベルの特徴を用いた研究が行われてきた[7),11)]．とりわけ文献 5) や 8) で報告されている，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が行う多様な発話による誤りを正しく検出する必要しなければならない．さらに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[3)]，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ことが報告されている[4)]．つまりユーザ情報はそのような音声対話システムでは非常に有用な情報であると言える．

一方，対話システムという状況を生かして，システムとの対話の間に信頼できると推定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[10)]．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[1)]，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．このように，もし音声認識結果が対話の後に正しいとみなせる部分があるなら，それらを教師信号とした機械学習を導入することができる．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せずに，システムのモデルの性能を向上させる可能性を持っている．

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，発話の取捨選択精度の向上への新しい特徴として利用する．この推定された音声認識率と，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，を各ユーザごとに計算したうえで，誤りが多いことで知られているバージイン発話の音声認識の正否の予測に用いる．ここで用いるユーザごとの推定音声認識率とバージイン率は，人手による書き起こしなしで実行時に取得可能である．

†1 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
†2 School of Computer Science, Carnegie Mellon University

表 1　バージインの有無による音声認識率

ASR results	Correct	Incorrect	Total	Accuracy
NO BARGE_IN	16,694	3,612	20,306	(82.2%)
BARGE_IN	3,281	3,912	7,193	(45.6%)
Total	19,975	7,524	27,499	(72.6%)

2. 暗黙的な教師信号に基づく音声認識率の推定

2.1 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して考える．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目的である．ここでバージイン発話とは，ユーザがバージインしながら行った発話である．我々の収集したデータにおいては，バージイン発話の解釈結果に誤りが多い傾向が見られ，これは主に音声認識誤りに起因するものが多い．表 1 に，分析対象としたデータ中の，プロンプトが最後まで再生された場合 (NO BARGE_IN) とバージインがあった場合 (BARGE_IN) 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，一部でも誤りが含まれる場合は誤りとして計数している．この表より，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とがわかる．同様の傾向は，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くなる，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[6] 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．

このため，とりわけ初心者ユーザがバージインする際に起こりがちであるこのようなエラーを検出するのは困難である．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以外の情報が導入される必要がある．我々は以前に，各ユーザのバージイン率を使ってバージイン発話の誤りを予測する手法を開発した[4]．これは直感的には，ユーザがそのシステム，特にそのバージイン機能に，どの程度習熟しているかに対応する．具体的には，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ことを利用している．

さらに我々は，各ユーザによる音声認識率にも着目する．なおここではバージイン発話を含む全ての発話に対する音声認識率を考える．この音声認識率もシステムへの慣れを表す指標のひとつであると言える．つまり，習熟したユーザの音声認識率は高いという傾向[3] に対応する．一方で，バージイン率と音声認識率は全てのユーザで必ずしも同時に向上するわけではないことも確かめられている[3]．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[3]．そこで我々は，そのユーザのそれまでのバージイン率と全発話の音声認識率の両方を使って，多様なユーザの熟練の度合を表現し，ここではバージイン発話のエラー予測という尺度で評価する．

2.2 暗黙的な教師信号を用いた音声認識率の推定

バージイン発話のエラー予測を，実行時に得られるデータのみから行う際に，当該ユーザの音声認識率を推定する．この際に我々は，各ユーザの対話パターンから得られる情報を活用する．具体的には,「京都駅前からでよろしいですか？」のように，ユーザの肯定または否定応答を導くシステムの明示的確認の後に来る応答の内容を活用する．

この際以下の仮定を置く．まずシステムの明示的応答に対するユーザ応答の音声認識結果は，肯定／否定の場合とも正しく認識できていると仮定する．次に，ユーザが肯定応答を行った場合，それに対応するユーザ発話の音声認識結果も正しいと仮定する．さらに，システムの音声認識結果が正しくない時には，ユーザはしばしば否定せず，新たな発話を始めることが多いので，肯定応答に対応する発話以外はすべて誤りであったと仮定する．すなわち，以下の 2 種類の音声認識結果は正しいとみなすことになる．

(1)　全ての肯定応答とその前の確認発話の対象となっていた発話
(2)　全ての否定応答

上記以外のユーザ発話は全て誤りであったとみなす．これに基づき，あるユーザの推定音声認識率は，当該ユーザのそれまでの発話を使って以下で計算する．

$$(\text{Estimated ASR accuracy}) = \frac{2 \times (\#\text{affirmatives}) + (\#\text{negatives})}{(\#\text{all utterances})} \quad (1)$$

単純な計算例を図 1 に示す．U2 と U4 は肯定／否定なので正解，U3 は肯定応答の直前の確認発話の内容であるため正解，U1 はそれ以外で不正解とみなすことで，この場合の U4 の時点での推定音声認識率は 0.75 となる．

2.3 バージイン率と推定音声認識率を用いたエラー予測

対話中の各時点での，そのユーザのバージイン率と音声認識率を入力として，ロジスティック回帰により予測したラベルの正否の精度を調べる．

U1: 京大正門前から
S1: 京都駅からでよろしいですか？
U2: いいえ
S2: ご乗車になる停留所名を言ってください
U3: 京大正門前
S3: 京大正門前からでよろしいですか？
U4: はい

図 1 推定音声認識率の計算例のための対話例

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

ここでの x_1 と x_2 は，現発話までのバージイン率と，一発話前までの音声認識率である．バージイン率の経時的変化に対応するため，バージイン率には窓をかけて古い履歴を考慮しないようにして精度を算出する[4]．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た場合に相当する．窓幅がユーザの発話数よりも大きい場合は，単に当該ユーザのそれまでの全発話を使って予測精度を算出する．このため，一ユーザによる最大発話数 2838 発話を窓幅が越えた場合は，単純にそれ以前の発話の全てを平均として使っている場合と等価になる．

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．

3. 実験的検証

3.1 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[5] で収集したデータを用いた．このシステムは，ユーザの要求に対応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．本システムは，電話を通じて一般市民からアクセス可能であった．システムは誤った出力を行わないように，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安全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，つまりユーザがシステムプロンプトの途中に話し始めた発話であった．

表 2 ASR accuracy by response type

	Correct	Incorrect	Total	(Acc.)
Affirmative	9,055	246	9,301	(97.4%)
Negative	2,006	289	2,295	(87.4%)
Other	8,914	7,009	15,923	(57.9%)
Total	19,975	7,544	27,519	(72.6%)

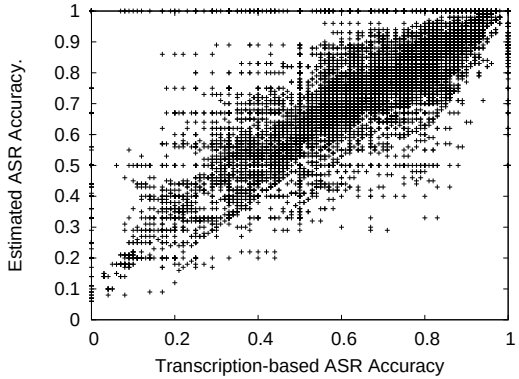

図 2 Correlation between transcription-based and estimated ASR accuracy

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各ユーザと仮定して分析を進める．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，この過程は妥当であるといえる．

各発話は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．音声認識結果の正否は発話ごとに計算し，書き起こし中に含まれる全ての内容語が正しく認識結果にも含まれている場合に正解とした．つまり一つでも書き起こし中の内容語が抜けていたり誤認識されている場合は，その音声認識結果は誤りとみなした．

3.2 暗黙的な教師信号を得るための仮定の検証

まず，システムの明示的確認に対する肯定／否定応答が全て正しいという仮定を検証する．ユーザ発話を音声認識結果に基づき肯定応答，否定応答，それ以外に分類し，それぞれ

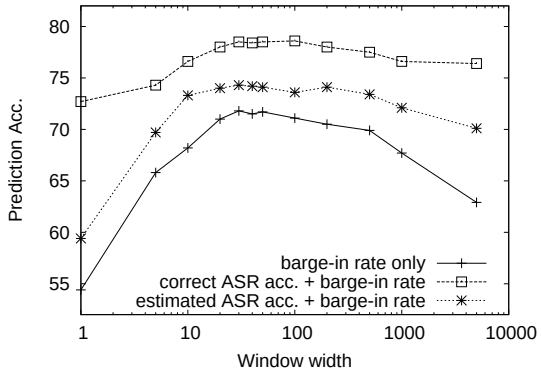

図 3 Prediction accuracy with various window widths

の音声認識精度，つまり適合率を計算したものを表2 に示す．ここでの肯定応答は「はい」の他に,「そうです」「OK」などを含み，否定応答は「いいえ」の他にも「違います」「だめ」なども含む．

表 2 に示されるように，肯定／否定応答の認識率は高い．これは「はい」や「いいえ」という発話が，語彙中の他のバス停や地名などよりも特に短く，区別しやすいことも一因であると考えられる．さらに，システムが「はい，またはいいえで答えてください.」というヘルプメッセージをしばしば提供していたのも肯定／否定応答の認識率が高かった一因であると考えられる．

次に，式 1 により推定した音声認識率と，書き起こしから計算した音声認識率との相関を調べた．図 2 の縦軸に 2.2 節で定義した推定音声認識率を，横軸に書き起こしに基づき計算した実際の音声認識率をプロットした．これらは，当該ユーザによる肯定／否定応答が一度以上行われた後の 26,231 発話全てに対して計算している．これらの間の相関係数は 0.806 であった．つまり，推定された音声認識率は正しい音声認識率と非常に高い相関を示していることが実験的に確認された．

3.3 バージイン率と推定音声認識率を用いたエラー予測の評価

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．この際，音声認識率は窓を設定しても精度が向上しないか悪化したため，窓の設定は行わず，当該発話までの全発話を使って各時点での音声認識率を計算した．この理由は，繰り返し使ったユーザの音声認識率は早々に収束するため[3]，バージイン率に比べて音声認識率に大きな変化がなかったためと考えられる．

表 3 Best prediction accuracies for each condition and window width w

Conditions (Used inputs)	Prediction acc. (%)
barge-in rate	71.8 (w=30)
correct ASR acc.	72.7
+ barge-in rate	78.6 (w=100)
estimated ASR acc.	59.4
+ barge-in rate	74.3 (w=30)

まず，音声認識率を使用する場合の効果を確認するために，書き起こしに基づく実際の音声認識率（図 3 と表 3 中では”correct”と表記）を，バージイン率とともにロジスティック回帰の入力に用いた．この場合，図 3 にあるように，エラー予測精度はバージイン率だけを使った場合に比べて大きく向上した．窓幅が 100 の時に最大精度 78.6%となっているが，窓幅が 30 からほぼ予測精度は収束している．またこの音声認識率だけを用いた場合の予測精度は表 3 中にあるように 72.7%で，バージイン率のみを使った場合の予測精度は 71.8%であった．これより両方を用いる場合の方が，いずれか一方のみを用いた場合よりも予測精度が高いことが示されている．これにより，バージイン率と音声認識率の両方が異なる情報を持っており，それぞれ予測精度の向上に貢献していることを確認した．

次に，上記の音声認識率を，2.2 節で述べた推定方法に置き換えた場合の精度を調べた．書き起こしに基づく実際の音声認識結果を用いた場合に比べて，全体に精度は劣化し，最大精度は窓幅が 30 の時の 74.3%であった．しかしこれは，バージイン率だけを用いた場合を上回っており，事後の人手による正解ラベル付与なしでエラーを予測する場合には，この推定音声認識率が有効であることが示された．

4. ま と め

本稿では，ユーザごとのモデルに基づく新たな特徴を用いて，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．本手法では，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．当該ユーザのそれまでの音声認識率は，発話のエラーを予測するうえでの有効な手がかりである．本稿ではそれを対話パターンから得られる特徴を用いて推定し，その有効性を評価実験を通して示した．

本手法は，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において有効である．今後の課題として，これ以外の対話戦略を採るシステムにおいて，推定音声認

識率の計算式やその実効性を検証する必要がある．また本手法で新たに得られた情報は，発話の信頼度計算における新たな特徴として利用できる．したがって音声認識の信頼度など，ボトムアップな情報と統合した場合の精度の検証も今後の課題として挙げられる．さらに，現状では，肯定／否定応答を音声認識結果のまま単純に全て正解としている．この部分をより精密にすることで，音声認識率の推定精度の向上が見込める．最後に，本稿では各ユーザの情報は，電話番号から得ている．このような各話者を同定する情報源がない場合に，音声や画像からの話者同定と組み合わせることができれば，電話以外のアプリケーション，例えばヒューマンロボットインタラクションなどにも適用することが可能である．

謝　　辞

京都市バス運行案内システムでのデータ収集にご尽力くださった京都大学学術情報メディアセンターの河原達也教授に感謝します．

参 考 文 献

1) Bohus, D. and Rudnicky, A.: Implicitly-supervised Learning in Spoken Language Interfaces: an Application to the Confidence Annotation Problem, *Proc. SIGdial Workshop on Discourse and Dialogue*, pp.256–264 (2007).
2) Komatani, K. and Kawahara, T.: Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output, *Proc. Int'l Conf. Computational Linguistics (COLING)*, pp.467–473 (2000).
3) 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).
4) 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).
5) 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).
6) 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. Int'l Conf. Spoken Language Processing (ICSLP)*, pp.173–176 (2004).
7) Litman, D. J., Walker, M. A. and Kearns, M. S.: Automatic Detection of Poor Speech Recognition at the Dialogue Level, *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.309–316 (1999).
8) 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. INTERSPEECH* (2006).
9) Rose, R.C. and Kim, H.K.: A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems, *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp.198–203 (2003).
10) Sudoh, K. and Nanano, M.: Post-Dialogue Confidence Scoring for Unsupervised Statistical Language Model Training, *Speech Communication*, Vol.45, pp.387–400 (2005).
11) 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. North American Chapter of Association for Computational Linguistics (NAACL)*, pp.210–217 (2000).