音声対話システムにおける バージイン発話の分類とそれに基づくエラー検出

中島 大一 [†1]　　駒谷 和範 [†2]　　佐藤 理史 [†2]

[†1] 名古屋大学 工学部 電気電子・情報工学科 [†2] 名古屋大学 大学院工学研究科

{taichi_n, komatani, ssato}@nuee.nagoya-u.ac.jp

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，その誤りは最大の問題である．音声認識誤りに起因するシステムの誤動作や，冗長な確認を防ぐために，誤りを誤りとして棄却する必要がある．本稿では，ある発話の音声認識結果が誤りであるかどうかの判定を，発話のエラー検出と呼ぶ．発話のエラー検出は従来，その発話の音声認識結果の信頼度に基づき行われる [3]．これに加えて，音声対話システムに特有の情報，つまり各ユーザ毎の音声認識率やシステム発話への割り込みを行う率（バージイン率）をプロファイルとして用い，エラー検出精度を向上させる研究も行われている [1]．

本稿では，バージイン発話を詳しく分析しその特徴を用いることで，発話のエラー検出精度のさらなる向上を図る．ここでバージイン発話とは，システム発話の最中に入力された発話とする．システム側で観測されるバージインは，システム発話の最中に何らかの音声入力があったことを指すが，これは必ずしもユーザが意図的に行った発話であるとは限らない．この結果，システム発話が全て再生された後に行われた発話と比べて，バージイン発話の音声認識率は低いという現象に繋がっている [1]．このため，バージイン発話で実際に起きている現象を明らかにする必要がある．

具体的にはまず，発話のタイミングに着目して，バージイン発話の分類を行う．発話のタイミングは，システム発話の開始時刻からユーザ発話の開始時刻までとする．この際，特にシステムの発話開始直後のバージイン発話にはユーザの意図どおりでないものが多いため，これを詳細に分類する．次に，この分析によって得られた，音声認識誤りを示す特徴を加えることで，発話のエラー検出精度の向上を示す．

2　発話タイミングに着目したバージイン発話の分類

2.1　対象としたシステム

本稿では，分析対象として，京都市バス運行情報案内システム [2] で収集されたデータを用いる．このシステムでは，ユーザは電話を通じて音声で乗車場所と降車場所または系統番号を発話する．この発話に対してシステムは音声認識を行い，指定されたバスがいくつ手前まで接近しているかを音声で出力する．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つが埋められると，バスの接近情報が出力される．

システムの内部状態は以下の3つに大別できる．それぞれの状態を典型的なシステム発話とともに示す．

条件入力待ち　「ご利用になるバス停または系統番号をどうぞ.」

確認　「京都駅前からでよろしいですか？」

結果出力　「17系統のバスは2つ手前の銀閣寺道を出発しました.」

ユーザ発話の音声認識結果にバス停名や系統番号などの内容語が含まれると，[条件入力待ち] から [確認] へと移行し，その内容語の確認が行われる．この [条件入力待ち] と [確認] のフローは，結果出力に必要な内容語が2つとも確認されるまで繰り返される．[確認] 状態では，はい，いいえなどの肯定否定表現のみを認識する言語モデルが用いられている．結果出力に必要な内容語が確認されると，[結果出力] に移行し，バスの接近情報を出力する．この後もう一度利用するかどうかを尋ね，利用する場合は [条件入力待ち] に戻る．

システムは次の2つのタイミングで応答を開始する．

1. 前のシステム応答から一定時間が経過した場合（**タイムアウト**）．この場合システムは同じ発話を繰り返し再発話を促す．

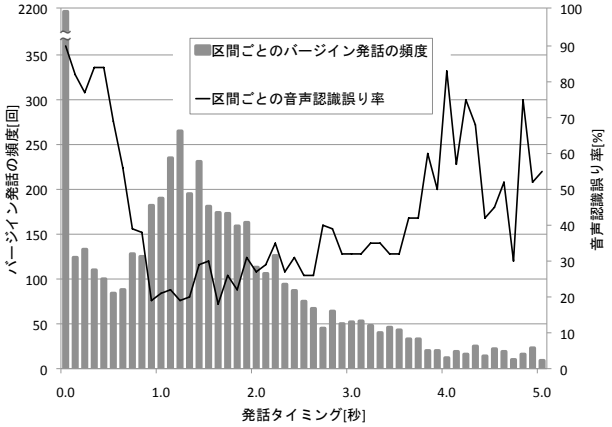

図 1: 発話タイミングごとのバージイン発話の頻度とそれらの音声認識誤り率

2. 入力がありそれを処理した直後．音声認識結果が得られた場合にはその内容に基づき応答し，入力が棄却された場合には再発話を促す．

システム発話の途中にユーザからの入力があり，システム発話が最後まで再生されなかった場合には，バージインが記録される．本稿では，バージインが記録された際の音声入力を，バージイン発話とする．

2.2 タイミングごとの発話の分布

まず発話タイミングに注目し，バージイン発話の特徴を分析する．ここで発話タイミングとは，システムの発話開始時刻から，それに対するユーザ発話の開始時刻までの時間とする．2002 年 5 月から 2005 年 3 月分のデータから得られたバージイン発話 7,193 発話を対象として，発話タイミングごとのバージイン発話の頻度とそれらの音声認識誤り率を調べた．音声認識誤りは，発話ごとに，人手による書き起こしに基づき判定した．つまり一発話中の内容語に一部でも誤りが含まれる場合，誤りとした．

結果を図 1 に示す．図の横軸は発話タイミングであり，0.1 秒区切りで集計している．縦軸は左側がバージイン発話の頻度，右側がそれらの音声認識誤り率である．バージイン発話の頻度を見ると，システム発話開始からおおよそ 2 秒以内にバージインが多く行われている．また音声認識誤り率を併せて考えると，システム発話の開始直後に，多くの音声認識誤りを含む発話が多かったことがわかる．

2.3 システム発話開始直後のバージイン発話の分類

前節の結果より，システム発話の開始直後に，多くの音声認識誤りが発生している．ここではより詳細に

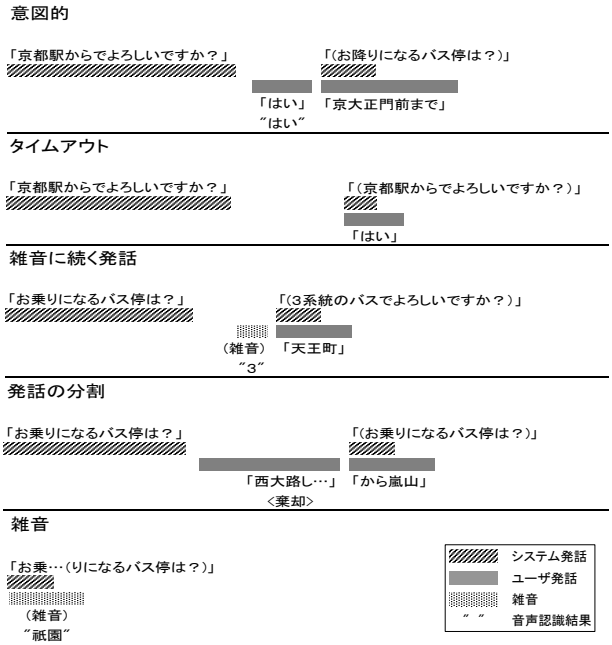

図 2: 発話の衝突の例

分析を行うために，2004 年 10 月 1 日から 20 日までの間の発話に対象を絞り議論を進める．本節では便宜上，発話タイミングが 0.8 秒以前のバージイン発話をシステム発話の開始直後のバージイン発話とみなし，この期間の全バージイン発話 119 発話を対象として分析を進める．これらは，システムの発話開始とユーザの発話開始がほぼ同時になった場合に相当する．本稿ではこのような状況を**発話の衝突**と呼ぶ．

上記の発話の衝突が起こった 119 発話を調査し，その具体的状況に応じて**意図的**，**タイムアウト**，**雑音に続く発話**，**発話の分割**，**雑音**の 5 つに分類した．以下で順に説明し，それぞれの例を図 2 に示す．

意図的 ユーザが対話の流れを既に知っており，システム発話をほとんど聞かずに発話を行った場合．

タイムアウト タイムアウト後のシステム発話とユーザ発話が衝突した場合．

雑音に続く発話 ユーザ発話の直前に雑音やユーザによるつぶやきがあった場合．これは背景雑音が多い場所での発話でしばしば発生する．この雑音やつぶやき部分に対するシステム応答の冒頭に，ユーザ発話が衝突する．

発話の分割 ユーザの一発話が，発話区間検出の誤りにより複数の発話に分割された場合．分割された発話の前半に対するシステムの応答に，後半のユーザ発話が衝突する．

雑音 継続的な背景雑音やユーザのつぶやきが，バージインとして誤って検出された場合．

雑音に続く発話や**発話の分割**では，何らかの入力に対

表 1: 発話衝突の分類結果

	発話数	誤り数	
意図的	10	4	(40%)
タイムアウト	20	10	(50%)
雑音に続く発話	18	17	(94%)
発話の分割	32	32	(100%)
雑音	39	38	(97%)
合計	119	101	(85%)

するシステムの応答が，それに続く後半のユーザ発話と衝突している．このため，前半部分の入力に対する音声認識結果に応じて，具体的な状況はさらに異なる．まず，前半部分の音声認識結果が棄却された場合で，その部分に内容語が含まれていない場合には，特に実害はない．内容語が含まれていた場合には，その部分を後で再入力する必要がある．図 2 での**発話の分割**の例がこの場合に相当する．

次に，前半部分で音声認識結果が得られた場合では，その認識結果によりシステムの状態が後半部分の発話時には移行しているため，発話の後半部分の解釈が意図通り行われないことがある．つまり例えば，［条件入力待ち］状態で発話の衝突が起こった場合，前半部分の音声認識結果により状態が［確認］に移行するため，発話の後半部分の音声認識は［確認］状態での言語モデルにより行われ，正しく認識されないという不具合が生じる．図 2 での**雑音に続く発話**の例がこの場合に相当する．この例では，後半の「天王町」という発話が入力された時点では，システムは［確認］状態に移行しており，言語モデルは肯定否定表現の認識用のものに変わっているため，この発話は正しく認識されない．

対象とした 119 発話において，上記の 5 分類の出現回数とそれらの音声認識誤り率を，表 1 に示す．全体では，85%にあたる 101 発話が音声認識誤りであった．これは図 1 で示された傾向と一致する．一方で，発話の衝突に該当する全ての発話が音声認識誤りであったわけではなく，**意図的**や**タイムアウト**等では，音声認識結果が正しい発話も多いことがわかる．

3 バージイン発話の分類結果を用いたエラー予測

前章の分析により得られた特徴を用いて，バージイン発話のエラー検出を行う．前章で明らかになった状況を以下に挙げる特徴として表す．

x_1: 発話タイミングが T 秒以前であるとき 1，そうでないとき 0.

x_2: システムが状態を変化させたとき 1，そうでないとき 0.

x_3: システムが［確認］状態にあるとき 1，そうでないときを 0.

これらを順に説明する．x_1 は，システム発話開始直後の発話であるかどうかを表す．図 1 に示したように，システム発話開始直後のバージイン発話には，音声認識誤りが多く含まれる．前章では便宜的に，発話タイミングが 0.8 秒以前の発話をシステム発話開始直後の発話として分類したが，ここでは閾値 T の値を変化させ，発話タイミングがそれより早いか遅いかを特徴として用いる．T の値は 0.0 秒から 1.6 秒の 0.2 秒区切りで設定した．値は次章にて実験的に求める．

x_2，x_3 は，システム発話直後のバージイン発話であっても，音声認識が成功する場合を考慮した特徴である．x_2 は，バージイン時に，システム状態が移行していたかを特徴としている．例えば，**タイムアウト**の場合，システム発話は単純に繰り返されるため，次のシステム発話が入力される時点ではシステム状態は変化していない．一方で，**雑音に続く発話**や**発話の分割**において，前半の発話が誤って認識された場合には，システムの状態が［確認］などに移行し，言語モデルが変更されるため，後半のバージイン発話は正しく認識されない．このような状況を表すために x_2 を用いる．

x_3 は，バージイン時にシステムが［確認］状態にあるかどうかを特徴としている．図 2 での**意図的**の例で示したように，ユーザが対話の流れを理解しているときには，発話の衝突が起きても，その時のバージイン発話が正しく認識される可能性はある．図 2 の例は，発話の衝突は［条件入力待ち］状態で起こっている．一方［確認］状態の発話で，確認内容を聞かずにユーザが肯定否定応答を行うのは困難であると考えられる．このように状態に応じて状況は異なるため，これを表す特徴として x_3 を設定する．

4 評価実験

バージイン発話 7,193 発話に対して，発話のエラー検出精度を調べた．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 (式 1) により発話のごとに音声認識結果の正解不正解を予測する．ここでは目的変数として正解に 1，誤りに 0 を割り当てる．なおパラメータ a_k，b は，評価データに対する 10-fold cross validation により推定する．

$$P = \frac{1}{1 + \exp(-(\sum_{k=1}^{n} a_k x_k + b))} \qquad (1)$$

説明変数 x_k には，4 章で示した特徴 x_1,x_2,x_3 を用いる．さらに，発話のエラー検出に従来から用いられ

表 2: 説明変数別の最高予測精度

説明変数	予測精度 (%)
(1) CM + 文献 [1] + 提案する特徴	93.5
(2) CM + 提案する特徴	93.0
(3) CM + 文献 [1]	92.2
(4) CM	90.8

文献 [1] の特徴：バージイン率，推定音声認識率

表 3: 提案する特徴 x_k 別の最高予測精度

説明変数	予測精度 (%)
(a) x_1,x_2,x_3	93.5
(b) x_1,x_3	93.5
(c) x_3	92.8
(d) x_1	92.7

る音声認識信頼度 (CM)[3] と文献 [1] から得られた以下の 2 つの特徴を用いる．

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推定音声認識率

表 2 に説明変数の組み合わせによる予測精度を示す．ここでは，「提案する特徴」として x_1,x_2,x_3 を全て用いた場合の最高予測精度を示す．なお，x_1 での T は，0.6，0.8 のとき最も精度が高かった．CM は，音声認識信頼度を表す．

表 2 の結果より，条件 (2) と条件 (4) を比較すると，発話のエラー検出に従来から用いられている音声認識信頼度 (CM) に加えて，提案する特徴を用いることで，精度は 2.2%向上している．発話のタイミングやそのときのシステムの状態といった，音声認識結果には依存しない，バージイン時に見られる特徴を用いることが，発話のエラー検出精度向上に有用であることを示している．また，条件 (1) と条件 (4) を比較すると，音声認識信頼度 (CM) に文献 [1] の特徴と提案する特徴を加えることで，精度は 2.7%向上している．これらから，音声認識信頼度 (CM) とは異なる情報源から得られる特徴が発話のエラー検出精度向上に有効であることが示されている．

さらに，提案する特徴 x_1,x_2,x_3 がそれぞれ予測精度に与える影響について考察した．表 3 は，x_1,x_2,x_3 をそれぞれ変えて，音声認識信頼度 (CM) と文献 [1] の特徴とともに発話のエラー予測を行った場合の最高予測精度である．

表 3 の結果より，条件 (b)〜条件 (d) を比較すると，特徴 x_1 と特徴 x_3 は，単独で用いるより，同時に用いほうが精度向上に有用であることが分かる．さらに，条件 (a) と条件 (b) を比較すると精度はほぼ同じであるが，特徴 x_2 はわずかながら精度向上に寄与していた．

5 おわりに

本稿では，発話のタイミングに着目し，バージイン発話の分類をおこなった．分類結果から得られた特徴を，音声認識信頼度と関連研究 [1] の特徴と共に用いることで発話のエラー検出精度が向上することを示した．

分類時において，システム発話開始直後のバージイン発話に意図的でない発話が多いことを示した．今後の課題としては，システム発話ごとのユーザが聞くべき最小時間の特徴を捉え，発話のエラー検出精度をさらに向上させることが挙げられる．

謝辞 本稿では京都市バス運行情報案内システムで収集されたデータを用いた．当該プロジェクトを主導された京都大学教授の河原達也先生に感謝します．

参考文献

[1] 駒谷和範, 奥乃博．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．言語処理学会第１５回年次大会講演論文集, pp. 218–221, 2010.

[2] 駒谷和範, 上野晋一, 河原達也, 奥乃博． 音声対話システムにおける適応的な応答生成を行うためのユーザモデル．電子情報通信学会論文誌. D-II, Vol. 87, No. 10, pp. 1921–1928, 2004.

[3] 駒谷和範, 河原達也. 音声認識結果の信頼度を用いた頑健な混合主導対話の実現法．情報処理学会研究報告. SLP-30-9, pp. 39–44, 2000.