1O2-OS-18-4

音声対話システムにおける対話的特徴を用いた対システム発話の判別

Detecting System-Directed Utterances using Dialogue-Level Features

平野 明[*1] 駒谷 和範[*1] 中野 幹生[*2]
Akira Hirano Kazunori Komatani Mikio Nakano

[*1]名古屋大学 大学院工学研究科 [*2]ホンダ・リサーチ・インスティチュート・ジャパン
Graduate School of Engineering, Nagoya University Honda Research Institute Japan Co., Ltd.

We have developed a method to determine whether a user utterance is directed at the system or not. A spoken dialogue system should not respond to audio inputs that are not directed at it (e.g., a user 's mutter), and it therefore needs to detect such inputs to avoid unsuitable responses. We classify the two cases by logistic regression based on a feature set including utterance timing, utterance length, and dialogue status. We conducted experiments using 5395 user utterances for both transcription and automatic speech recognition results. Results showed that the classification accuracy improved by 11.0 and 4.1 points, respectively.

1. はじめに

音声対話システムにおいて，システムによる誤った応答や不要な応答は，重大な問題のひとつである．このような応答は，ユーザ発話に対する音声認識結果が誤っていた場合以外にも，ユーザが意図しない音がシステムに入力された場合にも生じる．具体的には，ユーザの独り言やあいづちなどがこれにあたる．このような入力音による誤動作を防ぐために，入力音がシステムに向けられたものであるか否かの判別は重要である．この判別がないと，システムに向けられたのではない発話，例えば，ユーザが思わずつぶやいた独り言に対して，システムが「もう一度言ってください」と促すことになる．一般ユーザが音声対話システムを使う際には，システムの応答に対して無意識に独り言やあいづちを発してしまうことが少なからずあるため，このような判別は実用上重要である．

従来，このような誤動作回避は，入力音の解析により行われることが多い．最も簡便に実現可能な方法として，Julius に実装されている，一定の発話長より短い入力は雑音とみなして無視するオプションが利用できる [Lee 09]．また，音響的特徴に基づきユーザ発話と雑音等を判別し，後者を無視する手法も示されている [Lee 04]．音声認識結果の言語的特徴や音響的特徴，他話者の発話情報を用いて，システムに向けた発話を検出する研究も行われている [Yamagata 07]．一般的には，入力発話をシステムが扱うべきか否かの判別は，音声認識結果が正しいかどうかという観点から行われており，信頼度に基づく棄却手法が開発されてきた [Lane 05]．信頼度の算出に，対話システムの状態を用いた研究も行われている [Walker 00, Raymond 05]．これに対して本研究では，音声認識結果が正しいか否かではなく，そもそも入力音が，システムに向けられたものであるかどうかの判別を試みる．この判別において，音声対話システムと人間との対話では，対話の状態や発話タイミングが重要な要素となると考える．これらを特徴として組み入れ機械学習を行い，特徴が有効かどうかを実験的に検証する．

つまり本研究では，システムに向けた発話とそうでない発話に対して，**受諾**と**棄却**というラベルを与え，これらを判別する．それぞれの定義を以下に示す．

連絡先: 駒谷 和範，名古屋大学大学院工学研究科，名古屋市千種区不老町 C3-1(631)，komatani@nuee.nagoya-u.ac.jp

S1:	ドイツ西部のアーヘンにある大聖堂で
S2:	北部ヨーロッパでは最古のものです
U1:	あーきれいー
S3:	理解できませんでした。もう一度言ってください。

"S" と "U" はそれぞれ，システム，ユーザによる発話を表す．

図 1: システム向けでない発話が誤動作を引き起こす例

受諾 ユーザがシステムに向けて応答を求めて行う発話である．例えば，システムに対する質問や要求，システムへの応答がこれにあたる．この場合，理解結果が得られなかった場合でも，システムは「もう一度言ってください」のように，何らかの応答を行う必要がある．

棄却 独り言やあいづちなど，ユーザが意図的にシステムに向けて行ったのではない発話である．このような発話に対しては，システムは応答してはならず，入力を無視する必要がある．

図 1 は，システム向けでない発話が誤動作を引き起こす対話例を示している．この例では，S2 でのシステムの説明に対し，ユーザが U1 のように独り言をつぶやいている．これはシステムに向けられた発話ではない．このためその後 S3 でシステムがユーザにもう一度繰り返すよう促すのは不要な発話である．対システム発話の判別を行うことによって，このような不要なシステム応答を防止する．

対システム向け発話の検出は，対話における adressee 推定の一部として捉えることができる．Zuo らは，対話ではなく単独発話に対して，マルチモーダル情報を統合することによる対システム発話を行った [Zuo 10]．また Oviatt は，計算機と人の両方がいる場面において，システムへ向けた発話を行う際の，ユーザの発話の変化について調査している [Oviatt 07]．この問題に対して我々は，前の発話との時間関係や対話状態など，対話システム特有の特徴を用いて取り組む．

2. 対システム発話の判別に用いる特徴

システム向け発話の判別に，本研究ではロジスティック回帰を用い，対象とする発話に対して受諾か棄却かを判別する．ロ

発話の長さ	**発話内容**
x_1: 発話長	x_7: 返答
直前の発話との時間関係	x_8: 要求
x_2: インターバル	x_9: 中断要求
x_3: ユーザ発話の連続	x_{10}: フィラー
x_4: 包含されたバージイン	x_{11}: 内容語
x_5: バージインタイミング	**音声認識から得る特徴**
システム状態	x_{12}: 音響尤度差
x_6: システム状態	

表 1: 判別に用いた特徴

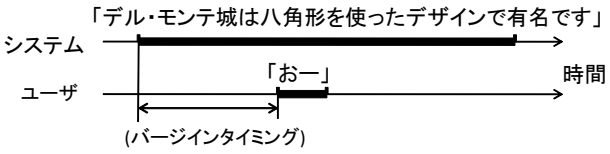

図 2: 発話区間がシステム発話に包含されるバージインの例

ジスティック回帰関数の目的変数として，受諾に 1，棄却に 0 を割り当てる．

$$P(x_1, \cdots, x_r) = \frac{1}{1 + \exp(-(a_0 + a_1 x_1 + \cdots + a_r x_r))} \quad (1)$$

x_k は以下で述べる各特徴の値，a_k $(1 \leq k \leq r)$ は説明変数である各特徴 x_k の係数であり，a_0 は定数項である．特徴の一覧を表 1 に挙げ，以下で順に述べる．特徴の設計に際して，実際の対話中で利用可能とするため，その発話までに得られる情報のみを使用した．各特徴の値は，値を算出した後，平均が 0，分散が 1 となるように正規化した．

発話の長さ

x_1 は入力された発話の長さを表す．単位は秒とする．これは，発話が長いほどユーザが意図して行った発話であるという傾向を表す．音声認識エンジン Julius でも，入力された音声の長さが一定値より短い場合，認識せずに棄却するオプション（-rejectshort）が実装されている．

直前の発話との時間関係

特徴 x_2 から x_5 は，直前の発話との時間関係を表す．x_2 は発話間間隔で，現在の発話の開始時刻と，その前のシステム発話の終了時刻との差と定義する．単位は秒とする．バージインが起こった場合，この値は負になるが，バージインは他の特徴で表現するため，x_2 は 0 とする．x_3 は，ユーザ発話が連続していることを表す．つまり，直前の発話がユーザによるものだった場合 1 とする．なお，一発話は，3.1 節で後述するように，ほぼ機械的に一定長の無音区間で区切ることで認定しているため，このようにユーザ発話やシステム発話が連続することはしばしば起こる．

バージインは，システムの発話中に，ユーザが割り込んで話し始める現象である．x_4 は，バージインのうち，ユーザの発話区間が，システムの発話区間に包含されている場合に 1 とする．つまり，ユーザがシステムの発話中に割り込んだが，システムより先に発話を止めた場合である．$x_4 = 1$ となる例を，図 2 に示す．一方，システム発話を，ユーザが発話により遮る，一般的なバージインの例を，図 3 に示す．この場合 $x_4 = 0$ となる．x_5 はバージインタイミングである．システム発話の長さに対する，システム発話の開始時刻からユーザ発話の開始時刻の間の時間の比を特徴とする．つまり，システムの発話開始時刻を 0，終了時刻を 1 として，システム発話のどの部分でユーザが割り込んだかを表している．システム発話長で正規化

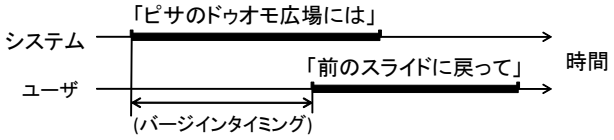

図 3: 一般的なバージインの例

81.66 - 82.32	S	すみません（保持）
83.01 - 84.13	S	わかりませんでした（保持）
84.81 - 88.78	S	もう一度、他の言い方で 質問してみてもらえますか？(譲与)
89.29 - 91.81	U	ギリシャの世界遺産について教えて下さい

“S” と “U” はそれぞれシステム，ユーザによる発話を表す．
“xx - yy” は，発話の開始，終了時刻（単位：秒）を表す．

図 4: ターンを譲与／保持するシステム発話の例

する前のバージインタイミングを，図 2 と図 3 に示している．

システムの状態

x_6 はシステム状態を表す．直前のシステム発話が，ターンを譲与するものである場合を 1，保持する場合を 0 とする．図 4 に例を示す．システムの発話のうち，1,2 発話目は，システムの応答に続きがあるため，ターンを保持していると考える．一方，3 発話目は，システムが話し終えてユーザに質問をしているため，発言権をユーザに譲与しているとする．この保持と譲与の認定は，システム発話に対して付与していた 14 種類のタグを分類することにより行った．

発話の言語表現

x_7 から x_{11} は，発話の言語表現中に，以下で挙げる表現が含まれることを表す特徴である．x_7 はシステムの発話に対する「はい」「いいえ」「そうです」など，返答を表す表現 11 種類が含まれるとき 1 とする．x_8 は,「教えてください」などの要求の表現 8 種類が含まれる場合を 1 とする．x_9 はシステムによる一連の説明を中断させる,「おわり」という単語が含まれる場合を 1 とする．x_{10} は，フィラーを表す「えーっと」や「へー」など表現が含まれる場合を 1 とする．フィラーは人手で 21 種類を用意した．x_{11} は内容語を表す 244 語のどれかが含まれる場合を 1，それ以外を 0 とする．内容語は，地域名や建物など，システムで使用される固有名詞である．

音声認識結果から得る特徴

x_{12} は，当該発話に対する音声認識結果と検証用音声認識器との間の，音響スコアの差である [Komatani 07]．検証用音声認識器の言語モデルには，Julius ディクテーション実行キットに含まれる，Web から学習した言語モデル（語彙サイズ 6 万）を使用した．この差を発話長で正規化したものを特徴とする．

3. 評価実験

3.1 対象データ

本研究では，音声対話システムを用いて収集した対話データ [Nakano 11] を対象とする．以下では，データ収集時の方法と，書き起こしの作成基準について説明する．ユーザは 19〜57 歳の一般男女 35 名（男性 17 名，女性 18 名）である．1 回 8 分の対話を，一人当たり計 4 回収録した．対話方法についてあらかじめ指示はせず，自由に対話するよう指示した．その結果，19415 発話（ユーザ：5395 発話，システム：14020 発話）を得た．収集した音声データを，400 ミリ秒の無音区間で機械的に区切って書き起こしを作成した．ただし，促音等，形態素内部では，400 ミリ秒以上の無音区間があっても，区切らず一

	受諾	棄却	計
書き起こし	4257	936	5193
音声認識結果	4096	202	4298

表 2: 実験対象の発話数

発話に含めた．400 ミリ秒より短いポーズは，当該部分に<p>を挿入して表記した．この発話ごとに，発話の内容を表すタグ 21 種類（要求，応答，独り言など）を，人手で付与した．

この書き起こしの単位と，受諾／棄却を判別すべきユーザ意図の単位は必ずしも合致しない．このため，短い無音区間を挟んで連続する発話を，マージして一発話とみなすという前処理を行う[*1]．この前処理は，書き起こしと音声認識結果それぞれに対して別に行った．書き起こしについては，ユーザの発話に対して付与したタグの中に，発話が複数に分かれていることを示すものがあるため，これが付与されている場合，二発話をマージして一発話とする．この結果，ユーザ発話数は 5193 発話となった．受諾または棄却の正解ラベルの付与は，これも人手で付与しておいたユーザ発話タグをもとに行い，受諾が 4257 発話，棄却が 936 発話となった．一方，音声認識結果に対しては，発話間の無音区間が 1100 ミリ秒以下のものをマージした．この結果，発話数は 4298 発話となった．音声認識結果に対する正解ラベルは，書き起こしと音声認識結果の時間的な対応関係に基づき付与した．具体的には，音声認識結果の発話開始または終了時刻が，書き起こしにおける発話の区間内にある場合，その音声認識結果と書き起こしデータ内の発話は対応するとする．その後，書き起こしデータにおける正解ラベルを，対応する音声認識結果に付与した．

実験に用いる発話数を表 2 にまとめる．音声認識結果の発話数が少ないのは，発話断片が前後の発話とマージされたことや，人手では書き起こされていた発話のうち，音声認識結果では発話区間が検出されなかったものが存在したためである．

3.2 実験条件

実験における評価基準は，受諾とすべき発話と棄却とすべき発話を，正しく判別できた精度とする．ロジスティック回帰の実装には，Weka[Hall 09] を用いた[*2]．式 1 中の係数 a_k は，10 分割交差検定により推定した．学習データ中で，受諾すべき発話数と棄却すべき発話数に偏りがあるため，棄却に対して，発話数の比に対応する重みを与え，学習と評価を行った．このため majority baseline は 50%である．

実験条件として，以下の 4 つを設定した．

1. 発話長のみを用いる場合
   特徴 x_1 のみで判別を行う．これは，音声認識エンジン Julius のオプション-rejectshort を用いる場合に相当し，簡便に実現できる方法であるため，ベースラインの一つとした．発話長のしきい値は，学習データに対して判別精度が最高となるよう定めた．具体的には，書き起こしに対しては 1.10 秒，音声認識結果に対しては 1.58 秒とし，それよりも発話が長い場合を受諾とした．
2. 全特徴を用いる場合
   表 1 に挙げた特徴を全て用いて判別を行う．書き起こしの場合は，音声認識から得る特徴 (x_{12}) 以外を全て用いる．
3. 音声対話システム特有の特徴を除いた場合
   上記の「全特徴を用いる場合」から，音声対話システム特有の特徴，つまり x_2 から x_6 を使用しない場合である．この条件を，もう一つのベースラインとする．
4. 特徴選択を行った場合
   利用可能な全特徴に対して，backward stepwise feature selection による特徴選択 [Kohavi 97] を行った場合である．つまり，特徴を一つずつ取り除いて判別精度を計算し，もし判別精度が悪化しない場合はその特徴を取り除く，という手順を，いずれの特徴を取り除いても判別精度が悪化するようになるまで繰り返した場合の結果である．

*1 ここでは，他の手法（例えば [Sato 02]）で発話の終了認定（end pointing）が正しく行えると仮定していることになる．

*2 具体的には “weka.classifiers.functions.Logistic” を使用した．

特徴選択を行った場合	85.4%
全特徴を用いた場合	85.1%
音声対話システム特有の特徴を除いた場合	84.2%
発話長のみを用いる場合	74.4%

表 3: 書き起こしに対する判別精度

特徴選択を行った場合	76.7%
全特徴を用いた場合	76.0%
音声対話システム特有の特徴を除いた場合	74.5%
発話長のみを用いる場合	72.6%

表 4: 音声認識結果に対する判別精度

3.3 書き起こしデータに対する判別性能

表 2 に記載されているユーザ発話 5193 発話 (受諾 4257，棄却 936) に対して，10 分割交差検定により判別精度を計算した．正解ラベルの偏りを考慮して，棄却とすべき発話に 4.55 (= 4257/936) の重みを与えて学習を行った．

4 つの実験条件に対する判別精度を表 3 に示す．全特徴を用いた場合の方が，音声対話システム特有の特徴を除いた場合よりも判別精度が高い．このことより，音声対話システム特有の特徴により，判別精度が向上したことがわかる．特徴選択では x_3 と x_5 が取り除かれた．発話長のみを用いるベースラインと特徴選択を行った場合を比較すると，判別精度は全体で 11.0 ポイント向上した．

3.4 音声認識結果に対する判別精度

ユーザ発話の音声認識結果 4298 個 (受諾 4096，棄却 202) に対して，同様に 10 分割交差検定による判別精度を計算した．音声認識には Julius を用いた．言語モデルの語彙サイズは 517 発話，音素正解率は 69.5%であった．正解ラベルの偏りを考慮して，棄却に 20.3 (= 4096/202) の重みを与えて学習を行った．

判別精度を表 4 に示す．やはり全特徴を用いた場合の方が，音声対話システム特有の特徴を除いた場合よりも判別精度が高い．この差は，マクネマー検定により統計的に有意であった．このことより，音声対話システム特有の特徴が，受諾と棄却の判別に有効であったことが示されている．特徴選択では，x_3，x_7，x_9，x_{10}，x_{12} の 5 つの特徴が取り除かれた．

特徴選択後の各特徴の係数について，表 5 にまとめている．係数 a_k が正であった特徴は，値が 1，もしくは大きいほど，その発話を受諾とする傾向を示している．負であった場合は，同様に，値が 1，もしくは大きいほど，棄却とする傾向を示す．例えば，x_5 の係数が正であったことから，バージインがシステム発話の後半に対するものであるほど，システムに向けた発話であったとする傾向が示されている．同様に，x_4 の係数が負であったことから，ユーザの発話区間がシステムの発話区間

係数 a_k が正	$x_1, x_5, x_6, x_8, x_{11}$
係数 a_k が負	x_2, x_4
特徴選択で除外	$x_3, x_7, x_9, x_{10}, x_{12}$

表 5: 音声認識結果に対する特徴選択後の係数

書き起こし： ニューヨークについて教えてください （判別結果： 受諾）
音声認識結果： 理由 よう 建築 が さ （判別結果： 棄却）

図 5: 書き起こしでは正解，音声認識結果では誤りとなる発話の例

に包含されていた場合には，システムに向けた発話ではなかったという傾向が示されている．

表 3 と表 4 を比べると，音声認識誤りにより，音声認識結果に対する判別精度は，書き起こしデータに対する判別精度よりも低い．また，音声認識結果に対する判別では，発話内容を示す特徴（x_7，x_9，x_{10}）が特徴選択の結果除外されている．これらの特徴は，音声認識結果に強く依存するため，音声認識誤りが多く生じた場合には有効ではなくなり，特徴選択により除外されている．

同一の発話に対して，書き起こしに対しては判別が正しく行われ，音声認識結果に対しては判別が誤りであった例を，図 5 に示す．この発話は，システムに対して要求を行うものであるため，正しい判別結果は受諾である．書き起こしの場合，「ニューヨーク」という内容語（x_{11}）と，「教えてください」という要求を表す表現（x_8）が得られており，これらの特徴が働くことで，正しく受諾と判別されている．一方，音声認識結果では，棄却すべき発話であると誤って判別されている．この例で典型的に示されるように，音声認識誤りが存在する場合には，発話内容に関する特徴は，正しく働くとは限らない．

図 6 に，音声対話システム特有の特徴を加えることによって判別が正解となった例を示している．図中のユーザ発話 U1 はフィラーであり，棄却と判別されるべき発話である．しかしその音声認識結果はアメリカという内容語を誤って含んでおり，音声対話システム特有の特徴を除いた条件では，誤って受諾と判別されていた．ここで，このユーザ発話 U1 から得られる特徴の一部を挙げる．システム発話 S2 の前半でユーザ発話 U1 が始まっているため，特徴 x_5 の値は小さい．またユーザ発話 U1 の発話区間は，システム発話 S2 の発話区間に包含されるため，特徴 x_4 の値も 1 となる．これらの特徴が働くことによって，音声対話システム特有の特徴を含む全特徴を用いた場合では，ユーザ発話 U1 は正しく棄却と判別された．音声対話システム特有の特徴は音声認識結果に依存しないため，音声認識結果が誤りがちである場合でも，発話の判別に有用である．

4. まとめ

本稿では，ユーザ発話が，システムに向けられたものであるか否かを判別する手法を提案した．判別に用いる特徴として，前発話との時間関係や対話の状態などの，音声対話システム特有の情報を用いた．これらの特徴を加えることで，発話長のみを用いるベースラインと比較して，受諾と棄却の判別率は，書き起こしで 11.4 ポイント，音声認識結果で 4.1 ポイント，それぞれ向上した．このことから，音声対話システム特有の特徴は，対システム発話の判別に有効であることを示した．

24.46 - 25.83	S1	ピサのドゥオモ広場には，
26.31 - 33.09	S2	ピサの斜塔をはじめとする，大理石で作られた壮麗な建造物群が立ち並んでいます。
26.66 - 28.44	U1	えーっと ［音声認識結果： アメリカと］
34.48 - 37.44	U2	えーさっき聞いたのでこれはいいです。

“S” と “U” はそれぞれシステム，ユーザによる発話を表す．
“xx - yy” は，発話の開始，終了時刻（単位：秒）を表す．

図 6: 音声対話システム特有の特徴を加えることによって判別が正解となった例

参考文献

[Hall 09] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H.: The WEKA data mining software: an update, *SIGKDD Explor. Newsl.*, Vol. 11, pp. 10–18 (2009)

[Kohavi 97] Kohavi, R. and John, G. H.: Wrappers for feature subset selection, *Artificial Intelligence*, Vol. 97, No. 1-2, pp. 273–324 (1997)

[Komatani 07] Komatani, K., Fukubayashi, Y., Ogata, T., and Okuno, H. G.: Introducing Utterance Verification in Spoken Dialogue System to Improve Dynamic Help Generation for Novice Users, in *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, pp. 202–205 (2007)

[Lane 05] Lane, I. and Kawahara, T.: Utterance verification incorporating in-domain confidence and discourse coherence measures., in *Proc. EUROSPEECH*, pp. 421–424 (2005)

[Lee 04] Lee, A., Nakamura, K., Nisimura, R., Saruwatari, H., and Shikano, K.: Noise Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs, in *Proc. ICSLP*, pp. 173–176 (2004)

[Lee 09] Lee, A. and Kawahara, T.: Recent Development of Open-Source Speech Recognition Engine Julius, in *Proc. APSIPA ASC*, pp. 131–137 (2009)

[Nakano 11] Nakano, M., Sato, S., Komatani, K., Matsuyama, K., Funakoshi, K., and Okuno, H. G.: A Two-Stage Domain Selection Framework for Extensible Multi-Domain Spoken Dialogue Systems, in *Proc. SIGDIAL Conference*, pp. 18–29 (2011)

[Oviatt 07] Oviatt, S.: Implicit user-adaptive system engagement in speech, pen and multimodal interfaces, in *Proc. IEEE ASRU*, pp. 496–501 (2007)

[Raymond 05] Raymond, C., Bechet, F., Camelin, N., De Mori, R., and Damnati, G.: Semantic Interpretation With Error Correction, in *Proc. IEEE-ICASSP*, Vol. 1, pp. 29–32 (2005)

[Sato 02] Sato, R., Higashinaka, R., Tamoto, M., Nakano, M., and Aikawa, K.: Learning decision trees to determine turn-taking by spoken dialogue systems., in *Proc. ICSLP* (2002)

[Walker 00] Walker, M., Wright, J., and Langkilde, I.: Using Natural Language Processing and Discourse Features to Identify Understanding Errors in a Spoken Dialogue System, in *In Proc. of ICML*, pp. 1111–1118 (2000)

[Yamagata 07] Yamagata, T., Sako, A., Takiguchi, T., and Ariki, Y.: System request detection in conversation based on acoustic and speaker alternation features, in *Proc. INTERSPEECH*, pp. 2789–2792 (2007)

[Zuo 10] Zuo, X., Iwahashi, N., Taguchi, R., Matsuda, S., Sugiura, K., Funakoshi, K., Nakano, M., and Oka, N.: Robot-directed speech detection using multimodal semantic confidence based on speech, image, and motion,, in *Proc. IEEE-ICASSP*, pp. 2458–2461 (2010)