Vol.2009-NL-191 No.14
Vol.2009-SLP-76 No.14
2009/5/22

バージイン発話タイミングモデルを導入した指示対象同定

松山匡子[†1] 駒谷和範[†1] 武田龍[†1]
尾形哲也[†1] 奥乃博[†1]

自然な会話を実現できる音声対話システムでは，ユーザが自由なタイミングや言語表現で発話できることが望ましい．我々は，ユーザが任意のタイミングでシステム発話に割り込み（バージイン）できる手法を開発している．本手法では，Independent Component Analysis (ICA) に基づくセミブラインド音源分離を利用している．本稿では，システムが列挙する項目に対してユーザがバージイン発話で指定した対象を同定するために，ユーザのバージイン発話から得られるタイミング情報を用いて解釈する新手法について報告する．まず，ユーザが参照表現を用いて発話する場合のタイミング分布を，予備調査の結果に基づき，ガンマ分布で近似する．次に，システムの読み上げる各項目に対して，ユーザ発話がそのタイミングで解釈されるべき場合とその音声認識結果で解釈されるべき場合とをそれぞれ確率として表現する．これら 2 つの確率を統合し，最も尤度の高い項目をユーザの指示対象と同定する．システムが列挙する項目の一つを指定するユーザのバージイン発話 400 発話に対して，本手法が 2 つのベースライン手法 (音声認識結果のみから指示対象を同定する手法，及び，ユーザの発話タイミングのみから指示対象を同定する手法) よりも高精度に同定できることを実験により確認した．

キーワード: 音声対話システム，バージイン，発話タイミング，指示対象同定，確率的統合解釈

Identification of User's Referent with Barge-in Timing Model

KYOKO MATSUYAMA,[†1] KAZUNORI KOMATANI,[†1]
RYU TAKEDA,[†1] TETSUYA OGATA[†1] and HIROSHI G. OKUNO[†1]

In conversational dialogue systems, the user prefers to speak at any time and to use natural expressions. We have developed an Independent Component Analysis (ICA) based semi-blind source separation method, which allows users to barge-in over system utterances at any time. We create a novel method from timing information derived from barge-in utterances to identify one item that a user indicates during system enumeration. First, we determine the timing distribution of user utterances containing referential expressions and then approximate it using gamma distribution. Second, we represent both the utterance timing and automatic speech recognition (ASR) results as probabilities of the desired selection from the system's enumeration. We then integrate these two probabilities to identify the item having the maximum likelihood of selection. Experimental results using 400 utterances indicated that our method outperformed two methods used as a baseline (one of ASR results only and one of utterance timing only) in identification accuracy.

Index terms: spoken dialogue system, barge-in, utterance timing, identification of user's referent, probabilistic integrated interpretation

1. はじめに

音声対話システムでは，ユーザは自由な言語表現を使えるだけでなく，任意のタイミングで発話できることが望ましい．特に，システムはユーザの割込み（バージイン）発話を許容できる必要がある．例えば，システムが検索結果などを列挙する際，ユーザはある項目を指定するために割り込んで発話できるべきである．しかし実環境下にあるロボットと音声で対話を行う場合，ヘッドセットのような接話型マイクを介した音声対話システムとは異なり，任意のタイミングでユーザに発話を許可するのは困難である. なぜならシステム発話が実環境を通じてマイクに回り込み, ユーザ発話の誤検出や誤認識が起こるためである. 武田らは，システム自身の発話やその反響による影響を抑制する音源分離手法を開発している[1)]．この Independent Component Analysis (ICA) に基づく手法を用いることで，人間どうしが行う自然な会話のように，ユーザのバージイン発話を許容することが可能となる．この場合システムはバージイン発話から，音声認識結果だけでなくユーザが話し始めたタイミング情報も得ることができる．これら二つの情報を利用することで，ユーザのバージインを許容しながら自然な会話ができる音声対話システム（Barge-In-Able Conversational Dialogue System; BIACDS）を実現することができる．

BIACDS の一例として，システムとユーザは図 1 のような対話を行う． 図 1 において，ユーザはシステムが “銀閣寺” と読み上げた時点でバージインを行っている．BIACDS では，セミブラインド音源分離によりシステム発話との混合音からユーザ発話 “それ” のみを分離し，バージインタイミング情報とともに指示対象同定を行うモジュールに送る．このサブシ

†1 京都大学大学院 情報学研究科 知能情報学専攻
Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University

Vol.2009-NL-191 No.14
Vol.2009-SLP-76 No.14
2009/5/22

User: おすすめのお寺を教えて．
System: 10 件候補があるので読み上げます．“金閣寺”，“銀閣寺*”，· · ·
User: それ!
System: 銀閣寺ですね．銀閣寺は最も有名なお寺の一つで · · ·
(*はユーザのバージイン時点を示す)

図 1 対話例 (1)
Fig. 1 Dialogue example (1)

ステムは，ユーザが “それ” と言って指定した指示対象を同定する．ユーザのバージイン発話のタイミングを用いることで，ユーザが “銀閣寺” を指していることを判定できる．

本論文では，システムが項目を列挙する状況において，ユーザの指示対象を同定する手法について報告する．ユーザは選択肢の中から一つの項目を指定する際，代名詞やその項目自体，または項目の略称を用いる．このような項目を列挙する対話システムは，次の 2 点で重要である．1 点目は，ユーザは指示対象をタイミング情報を用いて指定できる点である．音声認識率が低い実環境下では，バージインタイミングは音声認識結果に比べて頑健に検出でき，信頼できる場合が多い．2 点目は，このような対話は情報検索タスクの検索結果出力部で必須だからである．情報検索タスクは音声対話システムにおける有望なタスクのひとつであり, 現在 Google[⋆1]や Microsoft[2] でも開発が進められている．

ユーザの用いる言語表現は，“それ” のような参照表現のみに制限されるべきではないので，我々はユーザの発話がタイミングで解釈されるべき場合とその発話内容で解釈されるべき場合との両方の場合を扱う．つまり，数値であるバージインタイミングと，文字列である音声認識結果という異なる 2 つの情報を統合して解釈する．この場合，以下の 2 つの課題がある．

(1) ユーザの指示対象同定に用いるバージインタイミングのモデル化
(2) タイミングと音声認識結果の統合

課題 (1) への対処として，ユーザの発話内容と発話タイミングの関係を調査する．課題 (2) への対処として，我々はタイミング情報と音声認識結果をそれぞれ確率で表現し，両方の情報を考慮しながら尤度が最大となる解釈を採用する枠組みを構築する．

これまでバージインは，音声対話システムに関する研究の一環として多くの研究者らによって扱われてきたが，その主な課題はバージインの検出であった[3),4)]．彼らの目的は如何に素早く正確にバージインを検出するかであった．また，McTear はユーザ発話を認識するためにどのようにシステム発話を中断するかに焦点をあて[5)]，Ström はバージインが誤って検出された場合のシステムの挙動について報告している[6)]．このように，バージインタイミングをユーザの意図解釈に積極的に用いた研究はなされていない．本研究では，ユーザのバージインが正しく検出されたと仮定し，バージインという発語行為を利用したユーザ意図の新しい解釈手法について述べる．

⋆1 http://www.google.com/goog411/

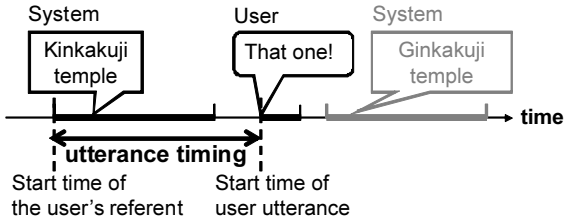


図 2 発話タイミングの定義
Fig. 2 Definition of utterance timing

2. ユーザの発話タイミングのモデル化

指示対象の同定にバージインタイミングを利用するために，ユーザの発話内容と発話タイミングの関係を調査した．本論文ではユーザの**発話タイミング**をユーザの発話開始時点と，ユーザが意図している指示対象をシステムが発話し始めた時点との差として定義する（図 2）．システムが項目を列挙し，ユーザがその中の一つの項目を指定する際，ユーザは**参照表現による発話**や**内容表現による発話**を用いる．前者を “それ” のように指示語を含む発話や，“今の” のようにタイミングを用いて指定する発話と定義する．後者を “金閣寺” のように内容語を含む発話と定義する．また，“二番目のニュース教えて” など番号で指定する発話，“足利義満が建てたほう” など，列挙項目に含まれる内容語が含まれていないが，列挙項目が発話内容から判断できる発話も後者に含む．ユーザが内容表現を用いる場合，発話タイミングは重要ではなく，その発話内容によって自分の意図を伝えている．一方ユーザが参照表現で意図を伝えようとする場合には，発話タイミングは重要であり，その分布には特徴があることが予想できる．

そこで，ユーザの参照表現の発話タイミングの分布を検証するために，表 1 の二つの異な

Vol.2009-NL-191 No.14
Vol.2009-SLP-76 No.14
2009/5/22

表 1 発話タイミングの調査条件
Table 1 Two different conditions for investigating utterance timing

	平均項目長 (秒)	ポーズ区間長 (秒)	発話数
条件 (1)	0.73	約 1.0	35
条件 (2)	5.27	2.0	69

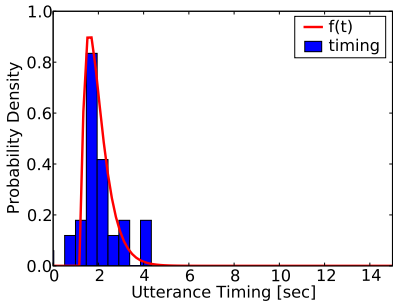


図 3 条件 (1) におけるタイミング分布
Fig. 3 Timing distribution in Cond. #1

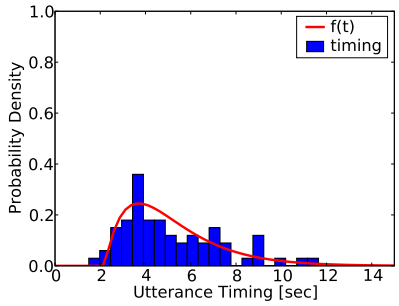


図 4 条件 (2) におけるタイミング分布
Fig. 4 Timing distribution in Cond. #2

る条件下でユーザの参照表現を収集した．平均項目長はシステムが列挙する項目の平均発話長であり，ポーズ区間長はシステムの列挙する項目間の時間差である．ユーザの発話タイミングは，図 2 に示すようにユーザの発話開始時刻を用いて算出する．ユーザの発話開始時刻は、分離されたユーザ発話を音声認識エンジン Julius[7] に入力したときの，Voice Activity Detection による発話の開始時刻とした．図 3，4 に，表 1 の二つの条件下で収集した発話タイミングの分布をヒストグラムで表す．ヒストグラムの幅は 0.5 秒である．ヒストグラムの高さは，その発話区間にある発話数を全発話数で正規化したものにヒストグラムの幅を乗じた結果を示している．これらの図から，発話タイミングの分布にはピークがあることがわかる．またそのピークの位置や減衰の度合は，平均項目長やポーズ区間長に応じてそれぞれ異なることがわかる．

Zhou らは，知覚の所要時間はガンマ分布に従うと示しており[8]，我々はこの知見に基づき, 参照表現の発話タイミングをガンマ分布でモデル化する．

$$f(t) = \frac{1}{(\rho-1)!\sigma^{\rho}}(t-\mu)^{\rho-1}e^{-(t-\mu)\frac{1}{\sigma}} \tag{1}$$

式 (1) のようにガンマ分布には，3 つのパラメータ μ，ρ，σ がある．σ は，分布の形状母数であり，発話タイミング分布にピークがあることを示す．本稿ではあらかじめ $\sigma = 2.0$ とする．また残りのパラメータはシステムが読み上げる一連の項目や，項目間のポーズ長に依存すると考えられるので，これらを基準として前もって値を決定する．まずパラメータ μ はシステムが発話する単語の平均長とする．μ は，システムがある項目を話し始める時点からユーザが発話する時点までのタイムラグに相当する．μ を単語の平均長とするのは，ユーザは項目を指定するまえに，一定時間のシステム発話を聞いて判断するからである．またパラメータ ρ はガンマ分布の減衰速度を表し，これをユーザが発話する前にシステムが読み上げた項目の長さの平均と，ポーズ区間の和に比例するとする．つまり，$\rho = \beta\times$ (平均項目長＋ポーズ区間長) とする．ここでは $\beta = 0.2$ とした．これらによりパラメータを決定したガンマ分布を図 3，4 に併せて赤線で示す．パラメータはそれぞれ，図 3 において $\mu = 1.2$, $\rho = 0.3$，図 4 においては $\mu = 2.2$，$\rho = 1.5$ となる．

3. バージインタイミングと音声認識結果を用いた指示対象の同定

本章では，発話タイミングと音声認識結果をそれぞれ確率で表現し，これらを統合して解釈する枠組について述べる．これにより，ユーザの指示対象を確率が最大となる項目として同定できる．

3.1 指示対象同定の枠組

確率 $P(T_i|U)$ を最大にするような T_i を求めることによって，指示対象同定問題を定式化する．ここで T_i はシステムが列挙する i 番目の項目であり，U はユーザ発話である．ユーザ発話 U は，発話タイミング t と音声認識結果 X の二つの要素を含むとする．つまり，$U = \{t, X\}$ とする．$P(T_i|U)$ は，システムが列挙する各項目に対して，ユーザ発話 U が項目 T_i を指示している確率を表す．すべての T_i に対する確率 $P(T_i|U)$ から，ユーザの意図した指示対象 T を求める．

$$\begin{aligned} T &= \underset{T_i}{\operatorname{argmax}}\, P(T_i|U) = \underset{T_i}{\operatorname{argmax}}\, \frac{P(T_i,U)}{P(U)} \\ &= \underset{T_i}{\operatorname{argmax}}\, P(T_i,U) \end{aligned} \tag{2}$$

式（2）より，実際は $P(T_i,U)$ を算出する．$P(T_i,U)$ は，次の二つの場合を考慮して計算

Vol.2009-NL-191 No.14
Vol.2009-SLP-76 No.14
2009/5/22

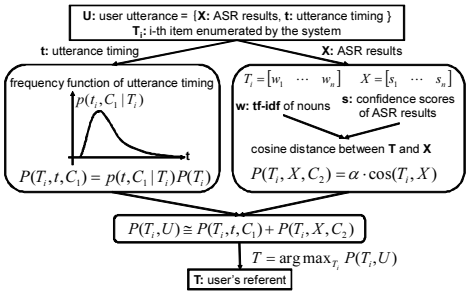


図 5 音声認識結果とタイミング情報を統合した指示対象同定手法の処理フロー
Fig. 5 Flow of identifying a user's referent

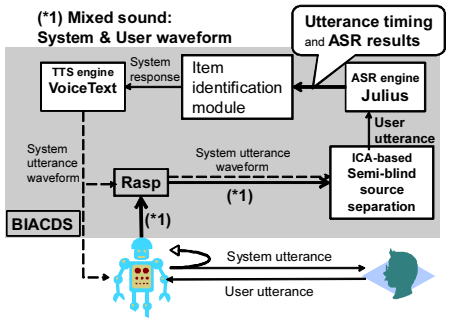


図 6 システム構成
Fig. 6 System architecture

される．ユーザがタイミングで意図を伝える場合と，音声認識結果で意図を伝える場合である．これらの場合をそれぞれ C_1，C_2 とする．$P(T_i, U)$ は次式で表される．

$$P(T_i, U) = P(T_i, U, C_1) + P(T_i, U, C_2) \tag{3}$$

式（3）に示すように，すべてのユーザの発話は，この二つの場合を考慮しながら解釈される．$P(T_i, U, C_k)$ は，あるユーザ発話 U が発話され，それが C_k として解釈される場合に，項目 T_i を指示している同時確率を表す．次節で順に $P(T_i, U, C_1)$，$P(T_i, U, C_2)$ について述べる．

3.2 発話タイミングを用いた指示対象同定

$P(T_i, U, C_1)$ の算出には，音声認識結果 X を用いずユーザの発話タイミング t のみを用いる．なぜなら C_1 はユーザが発話タイミングを用いて意図を伝える場合だからである．$P(T_i, U, C_1)$ は次式で算出される．

$$\begin{aligned} P(T_i, U, C_1) &\approx P(T_i, t_i, C_1) \\ &= P(t_i, C_1|T_i)P(T_i) \end{aligned} \tag{4}$$

t_i は，システムが列挙する各項目 T_i に対するユーザの発話タイミングを意味する．$P(t_i, C_1|T_i)$ は，ある項目 T_i に対し，ユーザがタイミング t_i で発話する確率を表している．この確率は 2 章で求めたガンマ分布に相当し，$P(t_i, C_1|T_i) = f(t_i)$ とする．すべての事前確率 $P(T_i)$ は等確率であると仮定し，$P(T_i) = 1/N$ とする．N は，システムが列挙する項目数である．

3.3 音声認識結果を用いた指示対象同定

C_2 の定義から，$P(T_i, U, C_2)$ は発話タイミング t を用いず，音声認識結果 X のみを用いて算出する．この確率は，ユーザ発話 U（すなわち音声認識結果 X）が，各項目 T_i にどれだけ近いかで表現する．この近さを，コサイン距離 $cos()$ を用いて表現する．

$$\begin{aligned} P(T_i, U, C_2) &\approx P(T_i, X, C_2) \\ &= \alpha \cdot cos(T_i, X) \end{aligned} \tag{5}$$

二つのベクトル T_i と X は，M 次元の要素からなる．M は，システムが列挙するすべての項目に含まれる名詞の総数である．

コサイン距離を計算するにあたり，各単語がその項目を表す重要度と音声認識誤りを考慮する必要がある．ある項目における各単語の重要度を示すために，ベクトル T_i の要素を TF-IDF 値[9] とした．IDF 値は，各列挙項目を一文書として計算した．音声認識結果が誤りである可能性を指示対象同定に反映させるために，ベクトル X の要素は，音声認識結果に含まれる各単語の信頼度とした．

係数 α は $P(T_i, U, C_1)$ と $P(T_i, U, C_2)$ のスコアレンジを調節するために用いる．本稿では $\alpha = 0.01$ とした．指示対象同定手法の全体の流れを図 5 に示す．

4. 評 価 実 験

4.1 BIACDS の実装

我々は図 6 に示すような構成で BIACDS を実装した．BIACDS の処理の流れは次のとおりである．まず，ユーザ発話とシステム発話がロボットに備え付けられたマイクに入力される．これらの混合音とシステム発話の音声波形を無線 RASP [*1]で同期させ，セミブラインド音源分離手法により混合音からユーザ発話を分離する．音声認識器 Julius[7] で分離されたユーザ発話を認識し，ユーザの発話開始時点を記録する．指示対象同定モジュールで音声認識結果と発話タイミングからユーザの指示対象を同定し，システムの応答を生成する．音声合成には VoiceText[*2]を用いた．

本システムは，自動的に RSS フィード上で更新されたニュースタイトルを取得し，読み上げる．さらに，ユーザのバージイン発話から指示対象を同定し，指示されたニュースの詳細をユーザに対して読み上げる．図 7 は，音声認識結果と発話タイミングの両方を考慮すべき対話例である．ユーザ発話 “留学生の記事について知りたい” は，音声認識結果のみからでは指示対象を同定できない．なぜならシステムが列挙する項目の両方に “留学生” という文字列が含まれているからである．この対話例のように，ユーザ発話が内容表現による

⋆1 Realtime Array Signal Processor (RASP). JEOL System Technology 社製の多チャンネル音響信号処理装置である．
⋆2 http://voice.pentax.jp/

Vol.2009-NL-191 No.14
Vol.2009-SLP-76 No.14
2009/5/22

System: “京都大学で**留学生**の受け入れ”,
“**留学生**イベント，きずな*で月例行事 · · · ”
User: **留学生**の記事について知りたい.
(*はユーザのバージイン時点を示す)

図 7 対話例 (2)
Fig. 7 Dialogue example (2)

発話である場合でも，音声認識結果と発話タイミングとの両方を解釈に用いるべきである．実際，本システムでは発話タイミングを用いることで，二番目の項目 “留学生イベント，きずな · · · ” をユーザの指示対象であると正しく同定できる．

4.2 実 験 条 件

評価用データとして，被験者 20 名から 400 発話を収集した．被験者には（1）システムは RSS フィードのニュースのタイトルを列挙するので，被験者が聞きたいものを指示すれば詳細が読み上げられるということ，(2）被験者は自分の好きなタイミングでシステム発話に割り込むことが可能で，項目を指定する際の言語表現は自由であることを教示した．システムが項目を列挙する場合の項目間のポーズ長は 1.5，2.0，3.0 秒の三種類とした．ガンマ分布のパラメータ μ は，あらかじめ 0.73 と設定した．各発話の後に，ユーザが実際に意図していた項目がどれであったかをユーザに確認し，同定実験における正解ラベルとした．

これらのデータに対し，指示対象の同定精度を算出する．つまり，ユーザが意図する指示対象が本手法より求めた指示対象と一致した率を算出する．比較のため，次の二つをベースラインとした．

ベースライン (1) 音声認識結果のみ

各ニュースタイトルと音声認識結果のコサイン距離からユーザの指示対象を同定する．コサイン距離が全て 0 の場合は結果は出力されず，同定失敗とする．

ベースライン (2) バージインタイミングのみ

ユーザが話し始めた時点の直前の項目をユーザの指示対象とみなす．

音声認識には CIAIR[10] の対話コーパスと RSS フィード中のタイトルを組み合わせた統計的言語モデルを用いた. 語彙サイズは 6,831 である. ベクトル X，T_i のサイズ M と列挙項目数 N は，列挙するニュースのタイトルの RSS フィード毎に異なる．平均して $M = 104.5$, $N = 15.8$ であった．ガンマ分布のパラメータ ρ は，2 章で述べたように，ユーザがバージインするまでに列挙した項目の平均発話長とポーズ区間長から求めた．

表 2 指示対象同定精度 (%) とそれぞれの場合における発話数 (#:)
Table 2 Identification accuracy [%] for user utterances (#: number of utterances)

	参照表現による発話 (#: 263)	内容表現による発話 (#: 137)	全発話 (#: 400)
(1) only ASR	4.2 (#: 11)	4.4 (#: 6)	4.3 (#: 17)
(2) only timing	84.8 (#: 223)	25.5 (#: 35)	64.5 (#: 258)
Our method	**88.2 (#: 232)**	**39.3 (#: 47)**	**69.5 (#: 279)**

4.3 実 験 結 果

収集した 400 発話のうちこのうち 137 発話は発話内容により指示対象が特定できる “内容表現による発話” であった．また 263 発話は発話内容からだけでは指示対象が特定できない “参照表現による発話” であった．“今のきずなのニュース教えて” 等，タイミングと発話内容の両方を用いて指示している発話については，発話内容のみからユーザの指示対象が特定できるため，“内容表現による発話” に分類した．これらの発話に対する単語正解精度は 35.8%であった．接話型マイクの代わりにロボットに備え付けられたマイクを使用したため，音源分離による歪みや音の反響が単語正解精度に影響していると考えられる．

本手法と二つのベースラインによる同定精度を表 2 に示す．音声認識結果のみを用いるベースライン（1）の同定精度は 4.3%であった．特に参照表現の同定精度は 4.2%と低かった．これは，参照表現は内容語を含まないため，音声認識結果のみからでは指示対象を同定できないからである．また内容表現における同定精度も低く，4.4%であった．これは接話型マイクを用いない音声認識が難しい状況下での単語正解精度の低さが原因である．

バージインタイミングのみを用いるベースライン（2）の同定精度は，64.5%であった．参照表現の同定精度は，ベースライン（1）に比べて 80.6 ポイント改善している．予想どおりではあるが，この結果からタイミング情報は列挙型の対話において有効なことがわかる．その上さらに，内容表現においてもベースライン（1）に比べて同定精度が 21.1 ポイント改善している．タイミング情報は内容表現による発話の解釈にも有効であることがわかる．

本手法の全発話に対する同定精度は 69.5%であり，二つのベースラインの精度を上回った．本手法とベースライン（2）の，参照表現による発話，内容表現による発話，全発話のそれぞれに対する同定精度の差は，有意水準 1%で統計的に有意であった．参照表現を用いた発話を含むすべての発話に対して，本手法の同定精度ががベースライン（2）より高いことは注目すべき点である．これにより，ユーザが発話タイミングにより意図を伝える場合であっても，音声認識結果を併せて解釈が有効であるといえる．

Vol.2009-NL-191 No.14
Vol.2009-SLP-76 No.14
2009/5/22

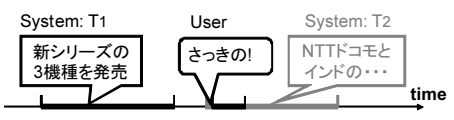
4.4 考　　察

ベースライン手法の両方で指示対象を同定できなかったが，本手法では正しく同定できた例を調査した．このような発話は，参照表現を用いた発話において 9 発話，内容表現を用いた発話において 8 発話存在した．図 8，9 にそれぞれ発話例を示す．図 8 の参照表現の場合，本手法では音声認識誤りにより指示対象でない項目の $P(T_i, U, C_2)$ が 0 より大きくても，発話タイミングによる解釈 $P(T_i, U, C_1)$ がより大きな値をとるため，指示対象を同定できた．図 9 の内容表現の場合，ユーザは発話内容により意図を伝えるのでタイミング情報はそれほど重要でないと考えられるが，列挙型の対話においては，参照表現と同様に指示対象の近くで発話することもある．このような場合，特に音声認識精度が低い状況下では，タイミング情報が指示対象同定に有効に作用した．図 8，9 の発話例について，ベースライン（1）では音声認識誤りにより T_1 以外の項目を指示対象とみなしたり，列挙項目と音声認識結果の距離を測れず同定に失敗していた．また，ベースライン（2）では単純なタイミングの解釈の結果 T_2 を指示対象とみなし，同定に失敗していた．本手法では音声認識結果とタイミング情報による解釈を統合することで，音声認識結果による解釈が曖昧な場合でも正しく指示対象を同定できた．

内容表現による発話のうちの 30 発話は，現状の本手法では正しく扱えない発話であった．例えば，“二番目のニュースを教えて”，“試合の結果を知りたいんだけど” などがこれらに含まれる．この場合，ユーザは発話内容により意図を伝えようとしているので音声認識結果による解釈が有効である．しかしこれらの発話は列挙項目に含まれる内容語を含まないため，単純にコサイン距離から音声認識結果と列挙項目との距離は測れない．今後の課題として，システムがこれらの発話を処理できるように実装することが挙げられる．前者の発話例に対しては，発話に含まれる番号と列挙番号を対応させればよい．後者に対しては，音声認識結果と列挙項目との潜在的距離を測るために，Latent Semantic Mapping[11] を用いるのが有効であると考えられる．

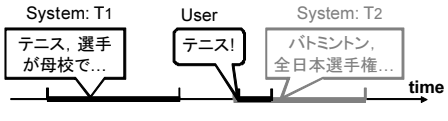
5. 結　　論

本稿では，ユーザのバージインタイミングをモデル化し，タイミングモデルと音声認識結果を確率的に表現し統合することで，ユーザの指示対象を同定する手法を開発した．また，RSS フィードから得られるニュース記事を読み上げる BIACDS を実装した．評価実験から，ユーザの 400 発話に対して本手法が音声認識結果やタイミング情報のみから解釈する場合よりも優れていることを示した．



ユーザはシステムがT2発話を話し始める時点でバージインする．
このとき，**“さっきの!”**が**“うなぎの.”**と誤認識され，T1以外の項目
Ti のP(Ti, X, C2) 値が0以上になる．
一方発話タイミングによる解釈では，P(T1, X,C1) が最も大きな値
をとり，これらの解釈を統合することでユーザの正しい指示対象T1
を同定できる．P(Ti, X,C2) は認識結果の信頼度に依存し，結果と
してタイミングによる解釈ほど重要視されない．

図 8　本手法で正しく同定できた参照表現の例
Fig. 8　Example referential expression correctly identified by our method



ユーザはシステムがT2発話を話し始める時点でバージインする．
このとき，**“テニス!”** が **“えっ．”** と誤認識され，すべての項目の
P(Ti, X,C2) 値が0になる．
一方発話タイミングによる解釈では，P(T1, X,C1) が最も大きな値
をとり，これらの解釈を統合することで，ユーザの正しい指示対象
T1を同定できる．内容表現の場合も，ユーザは指示対象付近で
発話することがあり，タイミングによる解釈が有効となる．

図 9　本手法で正しく同定できた内容表現の例
Fig. 9　Example content expression correctly identified by our method

本手法はシステムが選択肢を読み上げる中でユーザが一つを指定するという対話を対象とした．自然な会話によるインタラクションでは，ユーザは指示対象を示すためだけにバージインを行うわけではない．例えばユーザは，会話を早く終わらせたり，間違いを訂正したり，何か強く主張したいときに相手の発話に割り込む．本研究では，自然な会話ができる音声対話システムにおける直感的なインタラクション実現のための第一歩として，バージインタイミングを生かした新しいインタラクションを開発し，その結果ユーザの指示対象同定精度が本手法により向上することを示した．

参　考　文　献

1) Takeda, R., Nakadai, K., Komatani, K., Ogata, T. and Okuno, H.G.: Barge-in-able Robot Audition Based on ICA and Missing Feature Theory under Semi-Blind Situation, *Proc. IEEE/RSJ IROS*, pp.1718–1723 (2008).
2) Wang, Y.-Y., Yu, D., Ju, Y.-C. and Acero, A.: An Introduction to Voice Search, *IEEE Signal Processing Magazine*, pp.28–38 (2008).
3) Rose, R.C. and Kim, H.K.: A hybrid barge-in procedure for more reliable turn-taking in human-machine dialogue systems, *Proc. ASRU*, pp.198–203 (2003).
4) Ljolje, A. and Goffin, V.: Discriminative training of multi-state barge-in models, *Proc. ASRU*, pp.353–358 (2007).
5) McTear, M.F.: pSoken Dialogue Technology: Enabling the Conversational User Interface., *ACM Computing Surveys*, pp.90–169 (2002).
6) Ström, N. and Seneff, S.: Intelligent Barge-in in Conversational Systems, *Proc. ICSLP*,

Vol.2009-NL-191 No.14
Vol.2009-SLP-76 No.14
2009/5/22

Vol.2, pp.652–655 (2000).
7) Kawahara, T., Lee, A., Takeda, K., Itou, K. and Shikano, K.: Recent progress of open-source LVCSR Engine Julius and Japanese model repository, *Proc. ICSLP*, pp.3069–3072 (2004).
8) Zhou, Y., Gao, J., White, K., Merk, I. and Yao, K.: Perceptual Dominance Time Distributions in Multistable Visual Perception, *Biological Cybernetics*, Vol. 90, No. 4, pp. 256–263 (2004).
9) Salton, G.: *Automatic Text Processing*, Addison-Wesley (1989).
10) 河口信夫，松原茂樹，山口由紀子，武田一哉，板倉文忠：CIAIR 実走行車内音声データベース，電子情報通信学会技術研究報告，SP2003-136 (2003).
11) J.Bellegarda: Latent Semantic Mapping, *IEEE Signal Processing Magazine*, Vol.22, No.5, pp.70–80 (2005).