ロボット 音声対話におけるバージイン発話の指示対象同定

○松山 匡子, 駒谷 和範, 武田 龍, 高橋 徹, 尾形 哲也, 奥乃 博
京都大学大学院情報学研究科

Identification of User's Referent in Barge-in-able Robot Dialogue

*Kyoko Matsuyama, Kazunori Komatani, Ryu Takeda,
Toru Takahashi, Tetsuya Ogata and Hiroshi G. Okuno
Graduate School of Informatics, Kyoto University

Abstract—In conversational dialogue systems, users prefer to speak at any time. We allow users to barge-in over system utterances by utilizing an Independent Component Analysis based semi-blind source separation method. We create a novel method from timing-information to identify one item that a user indicates during system enumeration. First, we determine the timing distribution of referential utterances and approximate it by gamma distribution. Second, we integrate two probabilities, which are derived from utterance timing and automatic speech recognition results, to identify the item having the maximum likelihood. Experimental results using 400 utterances indicated that our method outperformed two baselines in identification accuracy.

Key Words: spoken dialogue system，barge-in，utterance timing，identification of user's referent

1. はじめに

人間とロボットの音声対話を実現するには, 周囲の雑音や部屋の残響への対処が必要である. なぜなら, ロボット自身にマイクが設置されているので, 目的音声以外の音もマイクに入力され, 音声認識が著しく困難になるからである. 音源分離/雑音下音声認識技術が発達した現在でも, 既存の音声対話システムを移植しただけでは, 実用に耐える性能を出すことは困難である. 従来のロボット研究で, 音声対話まで考慮して設計されたものはほとんど見られない.

本研究の目的は, バージイン発話, 特に**バージイン発話タイミング**に着目し, 実環境下でのロボット音声対話の実現である. バージインとはロボット発話へのユーザの割込み発話のことであり, 通常の音声対話システムでは頻発する. ロボットへの入力には, ロボット発話とユーザ発話が混合するが, Semi-Blind ICA といった手法により, それらは分離可能である. 特にバージインタイミングは, 分離歪みを含んだユーザ発話の認識と比較しても, 容易かつ頑健に検出ができ, 実際の対話を想定した有益な情報であるといえる.

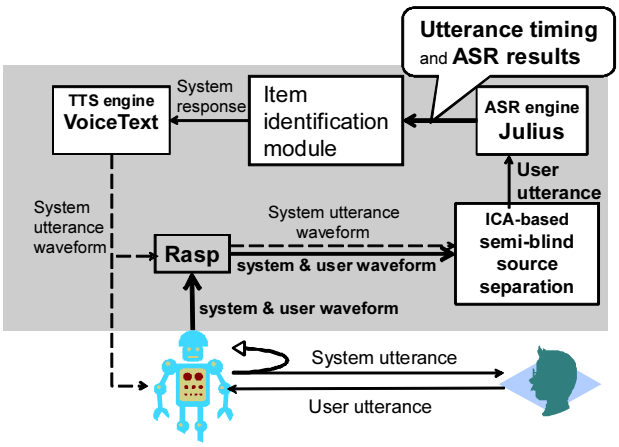

Fig.1 システム構成

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)

Fig.2 列挙型対話例

我々は列挙型対話に着目し, 音声認識結果とバージインタイミング情報の統合を行うことで, 実環境下でのユーザ発話の高精度な認識を実現する. ここでの課題は, 1) 統合方法の設計, 及び 2) タイミング情報のモデル化の 2 点である. 本稿では, 最尤推定法に基づいた確率的統合手法と, ガンマ分布を用いたバージインタイミングのモデリングを考案した. また本システムの評価は, 実際の音声対話実験で行う.

2. ロボット 音声対話システムの概要

本章で, 本研究におけるロボット音声対話システムを概説する. 図 1 に本システムにおけるデータフローを示す.

タスク: 列挙型音声対話　列挙型対話とは, ロボットが複数の選択肢を列挙し, その途中にユーザが 1 つを指定する対話である（図 2）. この対話は, (1) 情報検索タスクなどの検索結果出力部では必須で, 頑健な対話遂行が要求される, (2) ユーザの発話タイミングを生かした直感的なインタラクションが実現できる, という点から重要である.

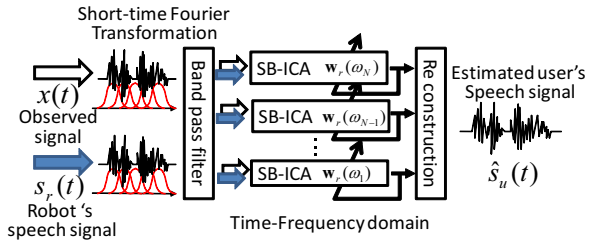

Fig.3 Semi-Blind ICA の処理概要

ICA-based Semi-Blind Source Separation　入力は，ユーザ発話とスピーカから出力されたロボット自身の発話の混合音，及び内部に保持したロボット発話である．これらは，無線 RASP [1] によって，同期され取得される．出力は，ロボット発話が分離された信号=ユーザ発話である．詳細は第 3 章で説明する．

ASR Engine: Julius　音声認識器 Julius [1] で，分離されたユーザ発話を認識し，認識された情報とユーザの発話タイミング情報が出力される．

Item Identification Module　音声認識結果と発話タイミングから，列挙型音声対話におけるユーザの指示対象を同定する．また，同定した結果から，現在の目的に沿った応答を出力する．**本稿での主題はこのモジュールの設計であり，第 4 章で詳細に説明する．**

Text To Speech Engine: Voice Text　生成された応答を実際に音声へ変換し出力する．音声合成には VoiceText [2]を用いた．

3. ICA-based Semi-Blind Source Separation

3·1 分離の仕組み

Semi-Blind ICA の分離に関して簡潔に説明する．このモジュールの入力は，時刻 t のマイク入力 $x(t)$ と内部に保持しているロボット発話 $s_r(t)$ である．ユーザ発話 $\hat{s}_u(t)$ は，次式に従って出力される．

$$\hat{s}_u(t) = x(t) - \sum_{n=0}^{K_r} w_r(n) s_r(t-n), \qquad (1)$$

ここで，w_r は未知のロボット発話の分離フィルタ，K_r はそのフィルタ長である．

既知音 $s_r(t)$ が仮想音源として観測されるとすると，この分離の入出力関係は正則な線形写像で記述できる．このため，通常の ICA を適用することで，ロボット発話の分離が可能である．本実装では w_r は各時間毎に逐次的に推定されため，$\hat{s}_u$ も逐次的に得られる．実際は，短時間周波数解析を行った後の，時間-周波数領域上で適用される（図 3）．詳細は文献 [2] を参照されたい．

3·2 手法の特徴

Semi-Blind ICA の特徴として，1) 雑音区間検出が不要なエコーキャンセラとして動作する，2) 時間-周波数領域で適用するため，演算量も実時間動作が可能な程度に抑えられている，点が挙げられる．また，ブラインド音源分離の技術がベースであるため，他の雑音源へ対応できるような拡張も可能であり，ロボットへの適用に適していると考える．

[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/

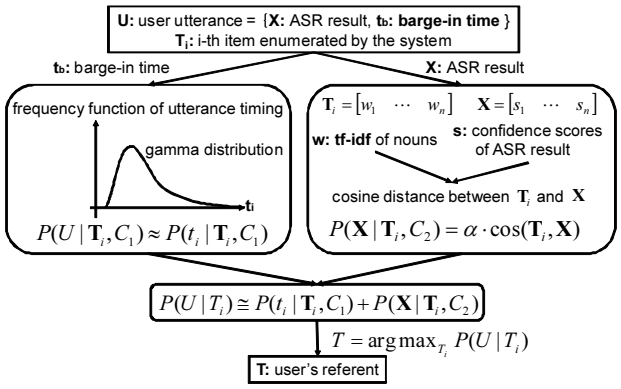

Fig.4 音声認識結果とタイミング情報を統合した指示対象同定手法の処理フロー

4. 発話タイミングを導入した指示対象同定

列挙型対話における指示対象同定モジュールについて説明する．本手法では，最尤推定に基づきバージインタイミングと音声認識情報を統合し，ユーザの指示対象を同定する．図 4 に処理フローを示す．

4·1 最尤推定による指示対象同定

今，利用できる情報は，ユーザ発話から得られる情報，バージインタイミング t_b と 音声認識結果 $\boldsymbol{X}$，の 2 つであり，これらを U と表す．我々の目的は，これらを用いて，ユーザが指示した確率が最も高い指示対象 $\boldsymbol{T}$ を求めることである．これは，最尤推定法では以下のように定式化される．

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

MAP 推定とは異なり，事前確率 $P(\boldsymbol{T}_i)$ は等確率であると仮定する．

次に，隠れ変数 c を用いて，「ユーザがタイミングで意図を伝える場合: $c = 1, C_1$」と「音声認識結果で意図を伝える場合: $c = 2, C_2$」を表現する．$P(U|\boldsymbol{T}_i)$ は変数 c に関する展開により，

$$P(U|\boldsymbol{T}_i) = \sum_k P(U|\boldsymbol{T}_i, C_k) P(C_k|\boldsymbol{T}_i) \qquad (3)$$

とできる．ユーザの意図 c の事前確率はまったく予想ができない，つまり，等確率であるとすると，次のように簡単化できる．

$$P(U|\boldsymbol{T}_i) = P(U|\boldsymbol{T}_i, C_1) + P(U|\boldsymbol{T}_i, C_2) \qquad (4)$$

式（4）に示すように，ユーザ発話 U はこの二つの場合を考慮して解釈される．$P(U|\boldsymbol{T}_i, C_k)$ は，「ユーザが項目 $\boldsymbol{T}_i$ を指示しており，かつ，C_k と解釈される場合に，ユーザ発話に関する情報 U が得られる確率」を表す．

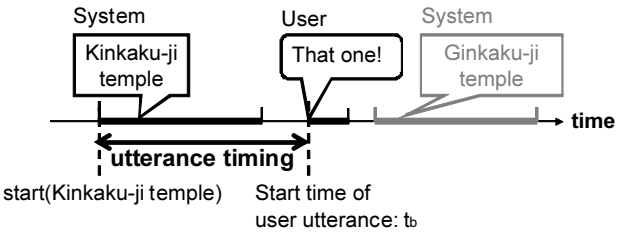

Fig.5 発話タイミングの定義

4·2 確率 $P(U|\boldsymbol{T}_i, C_1)$ のモデル化

本節では, ユーザのバージインタイミングを指示対象同定に利用するために, ユーザが参照表現発話を用いる場合の発話タイミングのモデル化を行う. 参照表現発話とは, ユーザが “それ”, “今の” のようにタイミングを用いて指定する発話であり, C_1 に相当する. ここで発話タイミング t_i を, システムの列挙項目ごとに設定する. つまり t_i を, システムの項目 $\boldsymbol{T}_i$ の発話開始時刻 $start(\boldsymbol{T}_i)$ とユーザの発話開始時刻 t_b との差, $t_i = start(T_i) - t_b$ と定義する (図 5).

C_1 の定義から, 我々は $P(U|\boldsymbol{T}_i, C_1)$ をユーザの発話タイミング t_i のみを用いてモデル化する. つまり,

$$P(U|\boldsymbol{T}_i, C_1) \approx P(t_i|\boldsymbol{T}_i, C_1) \tag{5}$$

が成り立つことを仮定する.

以降, $P(U|\boldsymbol{T}_i, C_1)$ をモデル化するために, 実際の発話タイミングを調査する. その後, 発話タイミングを近似する確率密度関数族について説明する.

4·2.1 発話タイミングの調査

ここでは, 二つの異なるロボット発話条件 (表 1) においてユーザの参照表現発話を収集し, 発話タイミングの分布を検証する. 平均項目長はロボットが列挙する項目の平均発話長であり, ポーズ区間長は列挙する項目間の時間差である. ユーザの発話開始時刻 t_b は, 分離されたユーザ発話を音声認識エンジン Julius [1] 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 6, 7 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. グラフの横軸は発話タイミング t_i であり, 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, 参照表現発話タイミングの分布にはピークが存在し, 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.

4·2.2 ガンマ分布を用いた発話タイミングモデル

Zhou らは知覚の所要時間はガンマ分布に従うと示している [3]. 我々はこの知見に基づき, 参照表現の発話タイミングを次式で示すガンマ分布でモデル化する.

$$P(t_i|\boldsymbol{T}_i, C_1) = \frac{1}{(\rho_i - 1)!\sigma_i^{\rho_i}}(t_i - \mu_i)^{\rho_i - 1} e^{-(t_i - \mu_i)\frac{1}{\sigma_i}} \tag{6}$$

3 つのパラメータ μ_i, ρ_i, σ_i は, ロボットが列挙す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ため, これらを考慮して設定する必要がある.

Table 1 発話タイミングの調査条件

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

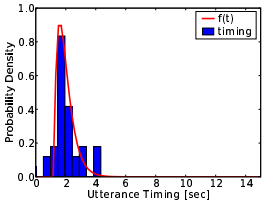

Fig.6 条件 (1) におけるタイミング分布

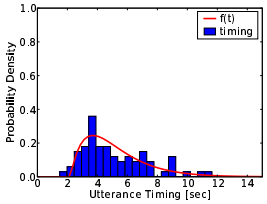

Fig.7 条件 (2) におけるタイミング分布

これらパラメータの設計については [4] を参照されたい. パラメータを決定したガンマ分布を図 6, 7 に併せて赤線で示す. パラメータはそれぞれ, 図 6 において $\rho_i = 2.0$, $\mu_i = 1.2$, $\rho_i = 0.3$, 図 7 においては $\rho_i = 2.0$, $\mu_i = 2.2$, $\rho_i = 1.5$ となる.

4·3 確率 $P(U|\boldsymbol{T}_i, C_2)$ のモデル化

C_2 の定義から, 我々は $P(U|\boldsymbol{T}_i, C_2)$ を音声認識結果 $\boldsymbol{X}$ のみを用いてモデル化する. つまり,

$$P(U|\boldsymbol{T}_i, C_2) \approx P(\boldsymbol{X}|\boldsymbol{T}_i, C_2) \tag{7}$$

を仮定する. この確率は, 「項目 $\boldsymbol{T}_i$ が与えられた下での, 音声認識結果 $\boldsymbol{X}$ が得られる確率」を意味する. これより, 音声認識結果 $\boldsymbol{X}$ と項目 $\boldsymbol{T}_i$ の特徴量を定義し, それらの距離に基づいた確率を設計する.

特徴量として, $\boldsymbol{X}$ は音声認識結果に含まれる各単語の信頼度, $\boldsymbol{T}_i$ は項目 i における各単語の重要度を示すために TF-IDF 値 [5] を用いた. IDF 値は各列挙項目 i を一文書として計算する. また, これらのベクトルのサイズは, システムのすべての列挙項目に含まれる名詞の総数 M とする. これらを特徴量として用いた理由は, 各単語がその項目を表す重要度を反映する必要があり, また, 音声認識誤りを考慮するためである.

これらのコサイン距離を確率 $P(\boldsymbol{X}|\boldsymbol{T}_i, C_2)$ の近似値として用いる.

$$P(\boldsymbol{X}|\boldsymbol{T}_i, C_2) \approx \alpha \cdot cos(\mathbf{T_i}, \mathbf{X}), \tag{8}$$

ここで, 係数 α は, 1) 式 (4) における確率間のスコアレンジを調節するため, 及び 2) 確率の正規化定数とするために設定している. 本稿では $\alpha = 0.01$ とした.

5. 評価実験

本実験では, 指示対象項目として, RSS フィードから得られるニュースタイトルを取り上げ, 本システムに実装した. ユーザが意図した指示対象の同定率によって, 本システムを評価する.

5·1 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) ロボットの RSS フィードの

ニュースタイトル列挙中に, 任意の項目を指定すればその詳細が説明されること, (2) 被験者は自由なタイミングでロボット発話に割り込むことができ, 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. 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. データ収集後, ユーザに実際に意図していた項目を確認し, 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. 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　ユーザの発話開始時点（ポーズ区間中の場合は直前）の項目をユーザの指示対象とみなす.

音声認識には CIAIR [6] 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 RSS フィード毎に異なり, 平均 5834.9 である. ベクトル $\mathbf{X}$, $\mathbf{T_i}$ 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 フィード毎に異なり, 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は, 4·2 節に従い決定した.

5·2 収集データ

収集した 400 発話のうち, 被験者がタイミングで意図を伝えた発話 (C_1 発話) は 263 発話であり, 音声認識結果で意図を伝えた発話 (C_2発話) は 137 発話であった. 後者は “留学生の話が知りたい” 等 RSS フィード中の名詞を用いた発話や, “二番目のニュース教えて” 等番号で指示対象を指定する発話を含む. “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, C_2 発話に分類した. 音声認識結果とユーザ発話の書き起こしに RSS フィード中の名詞や列挙番号が含まれる発話に対して, 単語正解精度は 41.8%であった. 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.

5·3 実験結果

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. ベースライン (1) における C_2 発話の同定精度が 4.4%と低い. 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. またベースライン (2) の C_2 発話において, 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ことからタイミング情報は, 音声認識結果により意図を伝え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, C_1 発話, C_2 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. C_1 発話を含むすべての発話に対して, 本手法の同定精度がベースライン（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た解釈が有効であるといえる.

Table 2 指示対象同定精度 (%)

	C_1 発話	C_2 発話	全発話
(1)	4.2	4.4	4.3
(2)	84.8	25.5	64.5
本手法	**88.2**	**39.3**	**69.5**

C_2 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名詞を含まないため, 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. 現在, システムがこれらの発話を処理できるよう実装中である. 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る. 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 [7] を用いることが有効であると考えられる.

6. おわりに

我々は, Semi-Blind ICA を用いたバージイン許容音声対話システムとして, RSS フィードから得られるニュース記事を読み上げる列挙型対話システムを実装した. またユーザの指示対象同定のために, 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. 評価実験から, ユーザの 400 発話に対して本手法による同定精度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, それぞれ 65.2 ポイント, 5.0 ポイント優れていることを示した.

参考文献

[1] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano. Recent progress of open-source LVCSR Engine Julius and Japanese model repository. In *Proc. ICSLP*, pp. 3069–3072, 2004.

[2] Ryu Takeda, Kazuhiro Nakadai, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Barge-in-able Robot Audition Based on ICA and Missing Feature Theory under Semi-Blind Situation. In *Proc. IEEE/RSJ IROS*, pp. 1718–1723, 2008.

[3] Y.H. Zhou, J.B. Gao, K.D. White, I. Merk, and K. Yao. Perceptual Dominance Time Distributions in Multistable Visual Perception. *Biological Cybernetics*, Vol. 90, No. 4, pp. 256–263, 2004.

[4] 松山匡子, 駒谷和範, 武田龍, 尾形哲也, 奥乃博. 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. 情報処理学会研究報告会, Vol.2009-SLP-76 No.14, 2009.

[5] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1988.

[6] 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.

[7] J.Bellegarda. Latent semantic mapping. *IEEE Signal Processing Magazine*, Vol. 22, No. 5, pp. 70–80, 2005.