実環境音声対話システムにおける バージイン発話タイミングを活用した指示対象の同定

松山 匡子 † 駒谷 和範 ‡ 白松 俊 ‡ 武田 龍 ‡ 尾形 哲也 ‡ 奥乃 博 ‡
† 京都大学 工学部情報学科 ‡ 京都大学大学院 情報学研究科 知能情報学専攻

1. バージイン発話が切り開く新しい地平

人間は自分の発話中に突然割り込みを受けても, その発話内容をうまく認識して応答を返すことができる. このようなバージイン発話が可能な音声対話システムを実現するための要素技術として, 実時間 **Semi-Blind ICA** 技術 [1] (以下, *SB-ICA* と略す) が最近開発されている. 本技術により, 常時マイク入力がオンになっていても, 自発話の回り込みによる誤動作を防止できるだけでなく, 自発話中にユーザが割り込んでも, ユーザのバージイン発話だけを抽出し, 音声認識を行うことができる.

SB-ICA からはバージイン発話の音声認識結果に加えて, ユーザ発話タイミング も得ることができる. すなわち, 音声認識結果とタイミング情報とを併用することで, 音声対話システムにおける新たなインタラクションが可能になると期待される. 例えば, ユーザが「それ」と言って指示したシステム発話中の対象を同定することが, タイミング情報を用いることで実現できる. この指示対象同定機能は, 実環境での頑健なユーザ発話理解に有効である. 本稿では, このようなインタラクションにおける, タイミング情報を用いたユーザの指示対象同定手法について報告する.

2. タイミングと音声認識結果を統合した解釈

本研究では, ユーザの発話表現を制限せず, 自由な発話でシステム発話中の対象の 1 つを指示させる場合を考える. この場合ユーザは, 意図する対象に含まれる内容語を発話したり, システム発話の最中に「それ」等の指示語で指示したりする. 特に, 意図する対象の表現が複雑であったり発音しにくい場合には, 指示語を用いる場合が増すと考えられる. これらのユーザ発話から指示対象を同定するために, システムは従来の音声対話システムのような音声認識結果に基づく解釈だけでなく, ユーザの発話タイミングを統合して解釈を行う必要がある.

我々は, 音声認識結果と発話タイミングによる対象の指示をそれぞれ確率で表現して統合し, ユーザの意図した指示対象を同定する. これにより, 音声認識結果や発話タイミングに曖昧性がある場合でも, 両方の情報から信頼できる部分を用いて総合的に最適な解釈が得られる.

3. タイミングのモデル化と確率による統合

ユーザ発話 U で指示された対象 T の同定を, 確率 $p(T_i|U)$ を最大にする T_i を求める問題として定式化する

Identifying User's Referents Using Barge-in Timing in Spoken Dialogue Systems under Real Environments: *Kyoko Matsuyama, Kazunori Komatani, Shun Shiramatsu, Ryu Takeda, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*

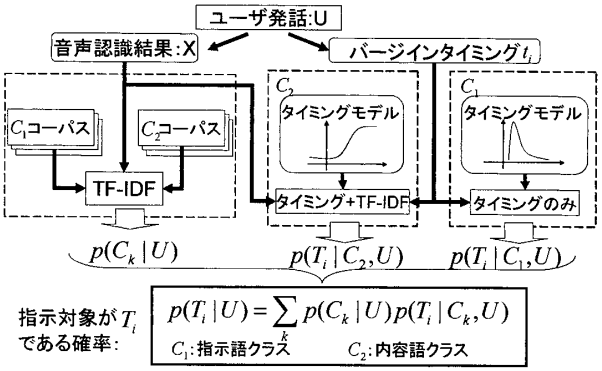

図 1: 音声認識結果とバージインタイミングの確率的統合

る. T_i はシステムが読み上げる i 番目の対象である.

$$T = \underset{T_i}{\operatorname{argmax}}\, p(T_i|U) = \underset{T_i}{\operatorname{argmax}} \sum_k p(C_k|U)p(T_i|C_k,U) \quad (k=1,2) \quad (1)$$

この $p(T_i|U)$ を, ユーザ発話の音声認識結果や発話タイミングから導かれる確率を用いて式 (1) で表現する. C_1 は指示語 (「それ」「いまの」「さっきの」等) を含む発話, C_2 は対象中の内容語を含む発話とする. 本手法の全体像を図 1 に示す.

以下, 3.1 節では指示語の発話タイミングから得られる確率 $p(T_i|C_1,U)$ について, 3.2 節で $p(C_k|U)$, $p(T_i|C_2,U)$ について順に説明する. なおユーザ発話 U は, その音声認識結果 X とシステムによる対象 T_i の読み上げ開始からの時間 t_i からなるとする. つまり $U=\{X,t_i\}$ である.

3.1 指示語発話のタイミングのモデル化

まず予備実験として, システムが対象 T_i を読み始めてからユーザが指示発話を行うまでの時間 t_i の分布を調べた. 被験者 10 名による指示語を含む 69 発話のタイミングを図 2 に示す. これからユーザの指示語発話はシステム発話開始から 3.5 秒付近に集中することがわかる.

この分布を確率密度関数として近似する. 知覚の所要時間に関する Zhou らの知見 [2] に基づき, 対象 T_i に対する指示語発話 (C_1) がタイミング t_i で生起する確率 $p(t_i|T_i,C_1)$ をガンマ分布により近似した.

$$p(t_i|T_i,C_1) = \frac{1}{(\rho-1)!\sigma^\rho}(t_i-\mu)^{\rho-1}e^{-(t_i-\mu)\frac{1}{\sigma}} \quad (2)$$

パラメータ値は $\sigma=1.5, \rho=2.0, \mu=2.2$ とした. 図 2 には $p(t_i|T_i,C_1)$ を赤色で併せて示した.

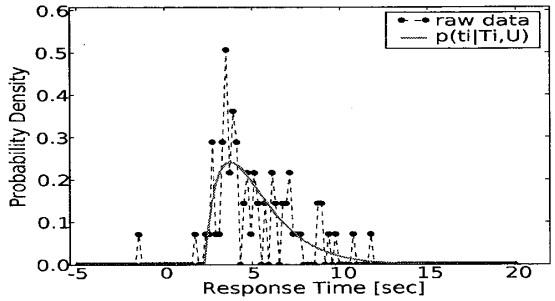

図 2: 指示語発話のタイミング分布

確率 $p(T_i|C_1,U)$ をこの $p(t_i|T_i,C_1)$ を用いて表す. 指示語発話 C_1 であれば音声認識結果は指示対象に無関係とし, さらに各対象の事前確率 $p(T_i|C_1)$ は等確率とすると, ベイズ則を用いて式 (3) が導かれる.

$$p(T_i|C_1,U) = p(T_i|C_1,t_i) = \frac{p(T_i|C_1)p(t_i|T_i,C_1)}{\sum_j p(T_j|C_1)p(t_j|T_j,C_1)} = \frac{p(t_i|T_i,C_1)}{\sum_j p(t_j|T_j,C_1)} \quad (3)$$

3.2 指示対象の確率的表現による統合解釈

式 (1) 中の $p(C_k|U)$, $p(T_i|C_2,U)$ を音声認識結果を用いて表現する.

$p(C_k|U)$ はユーザ発話が C_1,C_2 のいずれに属するかを表す. 音声認識結果が信頼できる単語のみを使うために, Julius[§]の単語信頼度 $C(w_i)$ を用いる. w_i は音声認識結果中の各単語 $(i=1,...,n)$ である.

$$p(C_k|U) = \frac{1}{n}\sum_i p(C_k|w_i)C(w_i) \quad (4)$$

$p(C_k|w_i)$ には C_1,C_2 に対応する語彙中の単語 w_i の TF-IDF 値を用いる. C_1 の語彙は対話コーパスで「それ」「これ」「いまの」「さっきの」を含む発話中の単語, C_2 の語彙はシステムが読み上げる対象に含まれる単語とした.

$p(T_i|C_2,U)$ は各対象の内容語を含む発話 U が各対象 T_i を指示する確率を表す. これを音声認識結果 X と各対象 T_i との距離 $\cos(T_i,X)$ と, 発話タイミング t_i によるペナルティを表す sigmoid 関数の 2 つを用いて定義する. 後者は, システムがまだ列挙していない対象をユーザが指示することは少ないという現象を表すために用いる. $\cos(T_i,X)$ は, 音声認識結果 X と各対象 T_i 中に現れる単語をそれぞれベクトルで表現した場合のコサイン距離である. 各対象 T_i に対応するベクトルの要素は T_i ごとの TF-IDF 値とした. sigmoid 関数では, まだ列挙されていない対象に対するペナルティを $b=0.99$ とし, 立ち上がり時刻を調整するパラメータを $m=2$ とした. 式 (5) の値を総和で正規化して $p(T_i|C_2,U)$ とする.

$$p^*(T_i|C_2,U) = \frac{b}{1+e^{-(t_i-m)}}\cdot\cos(T_i,X)+(1-b) \quad (5)$$

$$p(T_i|C_2,U) = \frac{p^*(T_i|C_2,U)}{\sum_j p^*(T_j|C_2,U)} \quad (6)$$

[§]http://julius.sourceforge.jp/

表 1: 指示対象の同定精度 (%)

手法	C_1 発話	C_2 発話	全発話
(1) 音声認識のみ	33	29	32
(2) タイミングのみ	93	10	67
本手法	**96**	**61**	**85**

4. 実装および評価実験

非接話マイクを用いて, 実環境を通じて音声対話を行うシステムを実装した. 本システムは RSS により読み込んだニュース記事のタイトルを列挙し, ユーザが指示したタイトルのニュース記事を読み上げる. ユーザはタイトルの列挙中にバージインすることもできる. ユーザ発話の分離には *SB-ICA* を用いた.

本システムを用いた評価データの収集法を以下に示す. 被験者にはバージイン可能なことを伝えた. タイトル間のポーズは 2 秒. 被験者 10 名より 100 発話を収集した. このうち C_1,C_2 に対応する発話はそれぞれ 69 発話, 31 発話である. 音声認識には CIAIR 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 (語彙サイズ 6,831) を用いた. 単語認識精度は 45.5%であった. 単語正解精度が低いのは, 非接話マイクを用いたためである. $p(C_k|U)$ 算出時に必要な対話コーパスには CIAIR [3] を用い, 457 文から C_1 用の語彙 442 語を得た. C_2 には RSS のタイトル 148 文から 1,084 語を用意した.

収集したデータに対して指示対象同定実験を行い, 本手法の精度を検証した. ベースラインは以下の 2 つとした. (1) 音声認識結果のみを用い, 各タイトル文とのコサイン距離により指示対象を決定する. 各タイトル文の語が音声認識結果に含まれない場合はランダムに選択する. (2) ユーザ発話開始時点のタイトルを指定対象とする. ポーズ区間で発話した場合は直前のタイトルとする.

指示対象同定実験の結果を表 1 に示す. タイミングを用いることにより, ユーザの指示対象の同定率の向上を示せた. C_2 において, 手法 (1) に比べて本手法の同定率が高いことから, C_1 だけでなく C_2 でもタイミングの導入が有効であることがわかる.

5. おわりに

本研究では, タイミング情報を活用したユーザ指示対象の同定手法について報告した. 今後は, 本手法をシステムへ実装し, 運用時のタスク達成までの対話からユーザビリティを評価する. それにより, バージインタイミングを用いた新たなインタラクションの可能性を検討する.

参考文献

[1] 武田他. 独立成分分析に基づく適応フィルタのロボット聴覚への応用. Vol. 26, No. 6, pp. 529–536, 2008.

[2] Y.H. Zhou *et al.* Perceptual dominance time distributions in multistable visual perception. *Biological Cybernetics*, Vol. 90, No. 4, pp. 256–263, 2004.

[3] 河口他. CIAIR 実走行車内音声データベース. 信学技報, SP2003-136, 2003.