情報処理学会論文誌　Vol. 52　No. 12　3374–3385 (Dec. 2011)

発語行為レベルの情報をユーザ発話の解釈に用いる音声対話システム

駒谷和範[†1]　松山匡子[†2]　武田龍[†2]
高橋徹[†2]　尾形哲也[†2]　奥乃博[†2]

本稿では，発話タイミングや発話後の沈黙のような，ユーザ発話の発語行為レベルの情報に着目して解釈を行う音声対話システムについて述べる．本研究では，実環境下でのヒューマノイドロボットとのインタラクションなど，音声認識が困難な状況での音声インタラクションの実現を目指している．具体的には，必要な場合に，音声認識結果に加えて，発話タイミングや発話後の沈黙などの情報を併用する部分対話へと対話を切り替える．この手法をシステムに実装し，31 名のユーザによる評価実験を行った．この結果，音声認識結果のみを用いる手法と比較して，音声認識率が低い状況でも，高いタスク達成率が得られることを確認した．

Spoken Dialogue System that Uses Information on Locutionary Acts to Interpret User Utterances

Kazunori Komatani,[†1] Kyoko Matsuyama,[†2]
Ryu Takeda,[†2] Toru Takahashi,[†2] Tetsuya Ogata[†2]
and Hiroshi G. Okuno[†2]

We constructed a spoken dialogue system that interprets user utterances by exploiting information on the locutionary act level such as the utterance timing and absence of an utterance. This study is useful for enabling spoken human-robot interaction under situations where automatic speech recognition (ASR) performance may be poor. In particular, our system can enter to dialogues when necessary in which the timing and absence of an utterance are used as well as the ASR results to interpret user utterances. We conducted an experiment with 31 participants. The result showed that our system achieved a higher task completion rate than a baseline system that uses only the ASR results when the ASR performance was not high.

1. はじめに

音声対話におけるユーザ発話の理解には，発話の内容（テキスト）の理解にとどまらず，発話という行為自体の包括的な理解が必要である．とりわけ実環境においてロボットと音声言語を用いて対話を行う場合では，周辺雑音や音源分離処理による音声認識精度の劣化は不可避であり，音声認識精度が低い場合を考慮に入れた対話の設計が必須である．音声対話システムや音声応答システムは，これまでにも多くの研究機関で研究が行われており[1)]，近年では一般ユーザに向けて公開されたものとして，京都市バス運行情報案内システム[2)]や，Let's Go! バスシステム[3)]，音声対話型京都案内システム[4)]，音声情報案内システム「たけまるくん」[5)]などがあげられる．これら従来の音声対話システムでは，音声認識結果として得られる単語列に基づいて応答を行うため，正しい音声認識結果が得られ難い環境ではなす術がない．ユーザの口元にマイクがあることを仮定しない状況でのヒューマノイドロボットとの音声インタラクションなどでは，音声認識性能の低下が著しく，つねに従来型の音声対話を行うのは困難である．さらに，実環境で一般ユーザに公開された音声対話システムでは，事前教示が行われないため，ユーザはしばしばシステムにとっての未知語を使用し，それが音声認識誤りを招く．このような実環境下での諸問題を打開する，新たなインタラクション技法が必要である．

本稿では，音声認識結果により得られる発話内容以外に，発語行為レベルの情報，ここでは発話のタイミングや発話後の沈黙を用いて，ユーザ発話を解釈し応答生成を行う手法について述べる．言語学者 Austin[6)] や Searle[7)] らは，発話行為理論（Speech Act Theory）を提唱し，発話をするという行為が，発語行為，発語内行為，発語媒介行為の 3 つの階層からなるとした．これらはそれぞれ，発話という行為そのもの，発話の内容を通じて行う依頼などの行為，発話の結果として聞き手に及ぼす行為，とされている．この中の発語内行為レベルの情報は，従来より対話システムではユーザ発話の意図として扱われており，当該発話に対する言語理解結果として利用されている．一方，発語行為には，発話の音韻的側面や統語的側面など，発話そのものの様々な側面が含まれる．本稿ではこのうち，声を発するという物理的な行為に着目し，これが発語行為の一部であると考えたうえで，発話の有無やその

†1 名古屋大学大学院工学研究科
Graduate School of Engineering, Nagoya University
†2 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

タイミングを発語行為レベルの情報として用いる．具体的には発話のタイミングと発話後の沈黙を扱う．実環境下では特に，声を発するという行為の検出は，正しい音声認識結果を得るよりも相対的に頑健に可能であると考え，これを音声対話システムでの発話の解釈に用いる．この情報を用いて，音声認識が困難な状況でもタスクを遂行できる状況を作り出すことを狙う．これら発話タイミングや沈黙の情報を取得するにはマイクを常時 ON にしておく必要があるが，ヒューマノイドロボット自身につけられたマイクを用いる場合，ユーザ発話の残響やロボット自身の発話がマイクに混入し，問題となる．本研究では Semi-Blind ICA 手法に基づく音源分離や残響除去[8] を行うことで，システム発話中のユーザの割込み（バージイン）や，そのような状況での無音状態や発話区間の検出を可能としている．

本稿では特に，実環境を模した実験環境において，発語行為レベルの情報を用いた解釈を音声対話システムが行うことで，対話全体の性能，具体的にはタスク達成率が向上するかどうかを検証する．このためにレストラン検索を行うシステムを構築し，一般から募集した被験者 31 名に用いて対話実験を行った．このシステムは，基本的にはユーザ主導で対話を行うが，必要な場合のみ，列挙型対話[9] や沈黙を用いた解釈を行う部分対話に移行する．列挙型対話とは，システムが選択肢を列挙し，その中の 1 つをユーザが指定するという部分対話のことを指す．この対話では，ユーザの発話タイミングを利用することで，音声認識結果に誤りが含まれる場合でも頑健に指示対象を同定できる．また，システムが持つ語彙をユーザに開示できるため，未知語を含む発話をユーザが繰り返すことによる対話の破綻を回避できるという副次的な利点もある．本稿ではまず，発話タイミングに関する関連研究について 2 章で述べた後，発語行為レベルの情報を用いた解釈の概要と列挙型対話への移行の判定方法について 3 章で述べる．次に 4 章では，実験を行うために構築したシステムについて述べる．システムのタスクは，データベース検索型のタスクである，京都市のレストラン検索とした．続いて 5 章で対話実験について述べる．この際，実際に音声認識精度が低い状況や未知語が存在する状況で，発語行為レベルの情報を用いた解釈が有効であったかどうかを検証する．具体的には，本手法を使用した際のタスク達成率や，音声認識精度とタスク達成率との関係について調査する．

2. 関 連 研 究

音声対話システムにおいて発話タイミングを扱う研究として，自然なターンテイキングを目指した研究がこれまでにも数多く行われてきた．発話交替やそのタイミングに関する研究は，下記に大別できる．以下ではまずこれらについて順に述べる．

(1)　音声対話システムにおける発話タイミングの分析
(2)　ユーザによるシステム発話への割込み（バージイン）の実現
(3)　適切なタイミングで相槌や応答生成などを行うシステムの構築

まず音声対話システムにおける発話タイミングを分析した研究をあげる．横山らは，うなずきなどの非言語情報の出現タイミングや，出力タイミングを変化させた場合の自然さなどについて分析した[10]．藤原らはタスク指向対話における人間同士の発話タイミングを，働きかけや応答などのタグごとに分析した[11]．さらに豊倉らは，人間同士の場合と，機械対人間の場合での差について，発話タイミングの違いについて分析した[12]．

次にバージインは，音声対話システムにおいてユーザビリティを高めるものとして認識されており，90 年代から実装が進められている[1]．そのうえで誤ってバージインを検出しシステム発話を停止してしまった際の戦略についての検討[13] や，バージイン発話をより早く検出したうえで，ユーザ発話の誤検出も防ぐ手法の開発などが行われている[14),15]．

音声対話システムにおけるタイミングのよい応答生成の研究としては以下があげられる．岡登らは句末の韻律情報をモデル化し，適切なタイミングで相槌を生成する手法を開発した[16]．また相槌のみに限らず，Nakano らは，ユーザの音声を逐次的に処理することにより，ユーザの発話の途中での相槌や応答生成が可能なシステムを構築した[17]．このシステムでは，ユーザの割込みに対してシステム発話を停止する機能も実現されており，柔軟なターンテイキング機構が実現されていた．Kitaoka らは，音響的特徴とキーワードを素性とし，人間同士の対話を学習データとした機械学習に基づき，適切な発話タイミングで相槌や応答を生成するシステムを構築した[18]．藤江らは同様に応答生成タイミングを決定し，またユーザの相槌も認識するシステムを構築している[19]．Raux らは，オートマトンモデルを用いて発話権の遷移を管理することにより，システムがユーザの発話中に誤って発話を開始することなく，素早い応答生成を実現している[20]．

本研究では，発話のタイミングや発話後の沈黙などの発語行為レベルの情報，つまり発話の有無から得られる情報が，音声認識結果よりも頑健に検出可能と考え，これらの情報をユーザ発話の解釈に積極的に用いる．これまでの研究との違いは以下の 2 点である．

(1)　ターンテイキングの実現ではなく，ユーザの意図の解釈に発話タイミングを用いる．
(2)　発話タイミングを，自然な対話に無意識に付随するものとしてではなく，音声認識が困難な状況下での意図伝達の手段として用いる．

このような情報を，ユーザの肯定・否定の意図の解釈に用いた研究としては以下があげられる．平沢らは，発話の継続時間長やシステム確認後のポーズ長などを特徴として用い，シ

ステム確認内容の正誤の検出を行った[21]．田中らは，ロボットが動作を学習する際の教示信号として，何も言わないという状態が肯定的な評価としてとらえられることを実験的に示した[22]．本研究では同様の情報を，タスク遂行型対話における意図的な伝達手段として利用する．

本研究ではさらに，データベース検索型のタスクにおいて，列挙型対話という新たな部分対話を新たに開発する．必要な場合にこの部分対話に切り替えることによって，音声認識が困難な環境下における頑健なタスク遂行を狙う．この列挙型対話では発話タイミングを意図伝達の手段として用いることができ，これが音声認識精度が低い状況下でのタスク達成率向上に役立つことを実験的に示す．

3. 発語行為レベルの情報を用いたユーザ意図解釈

3.1 発話タイミングを用いた解釈法

本章では，列挙型対話における，発話タイミング情報と音声認識結果を用いた指示対象同定手法について説明する．列挙型対話の概略と例を**図 1** に示す．図 1 の例のように，ユーザは内容語と指示語のどちらででも意図する対象を指示できる．また任意のタイミングで話してよい．つまり項目列挙中に割り込んで話すことも，項目をすべて聞き終わった後に発話することもできる．

ここでは指示対象同定問題を，$P(T_i|U)$ が最大となる T_i を求める問題として定式化する．T_i はシステムが列挙する i 番目の項目，U はユーザ発話を表す．このとき，事前確率 $P(T_i)$ は等確率とし[⋆1]，$P(U)$ は i に依存しないとすると，ユーザが意図した項目 T は以下の式で求まる．

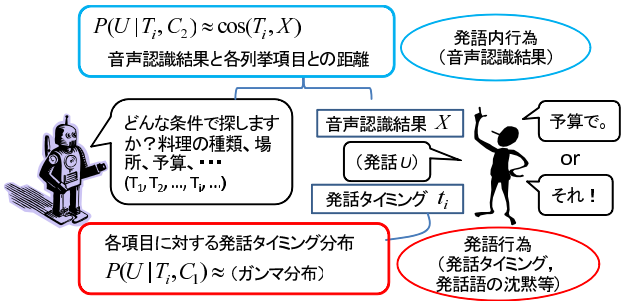

図 1　列挙型対話の例と概略
Fig. 1　Overview and example of enumeration dialogue.

$$T = \underset{T_i}{\operatorname{argmax}}\, P(T_i|U) = \underset{T_i}{\operatorname{argmax}}\, \frac{P(U|T_i)P(T_i)}{P(U)} = \underset{T_i}{\operatorname{argmax}}\, P(U|T_i). \tag{1}$$

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

$$P(U|T_i) = P(U|T_i, C_1)P(C_1|T_i) + P(U|T_i, C_2)P(C_2|T_i) \tag{2}$$

とできる．式 (2) に示すように，ユーザ発話 U はこの 2 つの場合を考慮して解釈される．$P(U|T_i, C_k)$ は，ユーザの意図する項目が T_i で，その発話を C_k として解釈する場合に，あるユーザ発話の特徴 U が得られる確率を表す．

ユーザ発話 U から得られる情報は，発話タイミング t_i と音声認識結果 X の 2 つであるとする．つまり $U = \{X, t_i\}$ とする．発話タイミング t_i は，システムが i 番目の項目の読み上げを開始した時刻と，ユーザの発話開始時刻の差と定義する．システムが列挙する項目が長い場合，ユーザは項目読み上げの途中で割り込む（バージインする）ことが多いため[23]，発話タイミングはユーザの発話開始時刻からの時間としている．タイミングの計算にユーザ発話の内容は依存せず，発話区間検出による時刻が用いられる．

T_i に対して，C_k の解釈の場合に U が発生する確率を $P(U|T_i, C_k)$ で表す．$P(U|T_i, C_1)$ は，ユーザの発話タイミングに対して仮定したガンマ分布を用いて計算する．C_1 の定義から，$P(U|T_i, C_1)$ をユーザの発話タイミング t [秒] を用いてモデル化する．つまり，

$$\begin{aligned} P(U|T_i, C_1) &\approx P(t|T_i, C_1) \\ &= \frac{1}{(\rho-1)!\sigma^\rho}(t-\mu_i)^{\rho-1}e^{-(t-\mu_i)\frac{1}{\sigma}} \end{aligned} \tag{3}$$

とする[9]．ユーザが項目 T_i の読み上げ時刻付近で発話した場合にはこの確率が高くなる．一方で，項目の読み上げをすべて聞き終えた後に発話した場合は，$P(U|T_i, C_1)$ はすべての T_i に対して低くなり，結果的に後述する音声認識結果に基づく解釈が行われる．パラメータ μ_i，ρ，σ はそれぞれ，システムが列挙している項目そのものの長さ [秒]，1.5，2.0 とし

⋆1 $P(T_i)$ には列挙内容に対するユーザの選好が本来反映されるべきである．ここでは $P(T_i)$ を学習する十分なデータがないため便宜的に等確率とした．

た[⋆1]．

$P(U|T_i, C_2)$ は，ユーザ発話の音声認識結果と，各項目 T_i をベクトル空間モデルで表現した際のコサイン距離を用いて計算する．C_2 の定義から，$P(U|T_i, C_2)$ を音声認識結果 X のみを用いてモデル化し，

$$P(U|T_i, C_2) \approx P(X|T_i, C_2) \approx \cos(\mathbf{T_i}, \mathbf{X}) \quad (4)$$

とする．ここでは，確率 $P(X|T_i, C_2)$ を，2 つのベクトル $\mathbf{X}$，$\mathbf{T_i}$ のコサイン距離で近似している．この 2 つのベクトルのサイズは，システムの列挙内容に含まれる名詞の総数 M である．このコサイン距離により，音声認識結果 X と各項目 T_i の近さを表す．ベクトル $\mathbf{X}$ と $\mathbf{T_i}$ の各要素にはそれぞれ，音声認識結果に含まれる各単語の信頼度，TF-IDF 値を用いた．TF-IDF 値は項目 i における各単語の重要度を示すために用い，IDF 値は各列挙項目 i を 1 文書と見なして計算する．音声認識結果 X が「それ！」のような指示語だった場合には，$\mathbf{T_i}$ に対応する要素が存在しないため，$\cos(\mathbf{T_i}, \mathbf{X})$ が 0 となり，結果としてタイミング情報のみに基づいて指示対象が決められる．なお今回の実験では $P(C_k|T_i)$ は等確率とし，タイミングによる解釈と音声認識結果による解釈が同等の重みを持つようにした[⋆2]．

3.2　発話タイミングを利用する部分対話への切替え

ユーザ主導形式で対話が進まない場合，つまり音声認識誤りによる対話の破綻のおそれがある場合に，列挙型対話へとユーザを誘導する．ここでは列挙型対話への切替え条件について述べる．切替えの判定は，対話の進行から得られる特徴と音声認識部から得られる情報を用いて行う．この判定はユーザの 1 発話ごとに行う．**表 1** に，判定に用いる特徴を示し，以下で順に説明する．

C_{same} は，システムが同一発話を繰り返す回数である．もし同一発話が繰り返されている場合，対話が進行していないことから，音声認識が困難な状況であると推定できる．ここでは，$C_{same} > 2$ である場合に列挙型対話へと移行する．

C_{listup} は，その対話における，それまでの列挙型対話への切替え回数である．これは，周辺環境自体が，音声認識が困難な状況かどうかの推定を試みている．つまり，それまでにすでに一定回数以上列挙型対話に移行していた場合，そのユーザの音声は認識されにくいか，周辺環境が雑音の大きい状況にあると見なす．ここでは，$C_{listup} > 2$ となった場合，以降の対話を列挙型対話で行う．

表 1　列挙型対話への移行判定に用いる特徴
Table 1　Features used when switching to enumeration dialogues.

特徴	判定条件
1. C_{same}: システムの同一発話回数	$C_{same} > 2$
2. C_{listup}: 列挙への切替え回数	$C_{listup} > 2$
3. V_{SNR}: SNR 推定値	$V_{SNR} <= 20$

V_{SNR} は，ユーザ発話ごとの信号対雑音比（SNR）推定値[24] である．この値が小さい場合は，ユーザ発話のパワーが小さいことを示し，これが誤認識の原因になる場合がある．ここでは，V_{SNR} の値がしきい値よりも小さい場合，まず 1 度ユーザにもう少し大きな声で話すように促し，それでも V_{SNR} の値がしきい値を下回る場合には列挙型対話に切り替える．このしきい値は，予備実験より $V_{SNR} = 20$ とした．

3.3　発話後の沈黙を用いた解釈

発話後の沈黙，つまり，ユーザがシステムの質問から数秒以内に発話したかどうかを用いて，ユーザの意図を推測する．具体的には，「○○でよろしいですか？」のように肯定か否定かを尋ねる質問の代わりに，「間違っていれば訂正し，正しければ 3 秒間そのままお待ち下さい」のように尋ねることで，ユーザが意図を伝えられるようにする．これにより，沈黙からユーザが肯定の意図を持つことが分かる．

本システムでは，確認へ応答する発話の音声認識が困難と見なされた場合，このような発話後の沈黙を用いる部分対話へと移行する．具体的には，システムの確認発話に対するユーザ発話が，一定回数以上繰り返して，肯定・否定・訂正を行う発話として解釈できない場合，発話後の沈黙による解釈にシステムを切り替える．なお訂正は検索条件となりうる内容語を含む発話を指す．ここでは，この繰返し回数を 2 回とした．また，肯定の意図を伝えるためにユーザが沈黙すべき時間は，実験時には 3 秒とした．

4.　実装した音声対話システム

前章で述べた手法の有効性を確認する実験を行うために，音声対話システムを実装した．本システムのタスクはデータベース検索で，ドメインは京都市のレストランを対象としている．対話例を**図 2** に示す．本システムの対話は大きく分けて以下の 3 つのフェーズからな

⋆1 σ はガンマ分布の尺度母数であり，発話タイミングの分布にピークがある場合は，タスクに依存せず $\sigma = 2.0$ 程度でよい．一方 μ_i や ρ は，システムが列挙する項目の長さや，項目間のポーズ長の設定に依存する．本稿での実験では，システムの列挙内容がほぼ 1 単語と短いことから，μ_i をそのときにシステムが列挙している項目の長さそのものとした．つまり結果的に，発話タイミングは各項目の読み上げ終了からの時間となっている．列挙内容が短いことから，ユーザが意図した項目に対する発話タイミングが負の値となることはほぼなかった．

⋆2 $P(C_k|T_i)$ の値を，列挙する項目長やユーザの特性に応じて変化させることによる，指示対象同定精度の向上も確認している[23] が，今回は簡単のため固定値とした．

——— (フェーズ I) ———
S1: 検索したいレストランの条件を言ってください.
U1: イタリア料理で京都駅周辺
S2: 検索条件は料理の種類がイタリア料理，場所が京都駅周辺で
よろしいですか？
U2: はい
S3: 該当するレストランは 6 件です．さらに条件を絞りこみますか？
U3: いいえ
——— (フェーズ II) ———
S4: 6 件のレストラン名を読み上げます.
詳細を知りたいレストラン名を指定してください.
リストランテキョウト, オイ ···
U4: それ！
S5: リストランテキョウトでよろしいですか？
U5: はい
——— (フェーズ III) ———
S6: リストランテキョウトについてどんな詳細情報が知りたいですか？
U6: （音声認識誤り）
S7: 理解できませんでした．もう一度お願いします.
···［繰り返し適切な理解結果が得られなかった］···
S9: 申し訳ありませんが認識しづらいようです.
選択肢を列挙しますので選んでください.
電話番号，営業時間，PR ポイント，···
U10: それ！
S11: PR ポイントですね．リストランテキョウトの PR ポイントは ···

図 2 対話例
Fig. 2 Example dialogue.

り，この順に対話を進める．

フェーズ I 検索条件の指定
システムは検索条件の入力を促し，ユーザは「料理の種類」「場所」「予算」から 1 つ以上を入力する．システムは該当するレストランの件数をユーザに示す.

フェーズ II 検索条件に該当するレストランの列挙
システムは検索条件に合致するレストランの名前を列挙する．ユーザはその中の 1 つを選択する.

フェーズ III 情報の提示要求
フェーズ II で選択されたレストランについて，ユーザは詳細情報（たとえば電話番号など）を尋ねる．システムはその内容を答える.

表 2 列挙型対話で列挙される内容
Table 2 Contents enumerated by system.

	列挙内容	列挙される語の例	
I	スロット名	料理の種類，場所，予算	
	料理の種類	和食，洋食，···	寿司，そば，···
	場所	左京区，上京区，···	出町柳，銀閣寺，···
	予算	5,000 円以上 10,000 円以下，···	5,000 円，6,000 円，···
II	レストラン名	（検索結果の上位 10 件）	
III	詳細情報の項目名	電話番号，営業時間，お店の PR ポイントなど 6 項目	

図 2 中にはフェーズの区切りも示している．ここでのフェーズ I と III は，一般的なデータベース検索における対話の進行に関するモデル[25] の 2 つのモードに相当する．また対話フェーズ I と III では，通常では図 2 中のフェーズ I の例のようにユーザ主導で対話が行われる．一方，図 2 中のフェーズ III のように，3.2 節で述べた条件が満たされた場合には，タイミングを用いて解釈を行う列挙型対話へと移行する．なおフェーズ II は検索結果の出力部分であるため状況にかかわらず列挙が行われる.

列挙型対話に移行した場合に，システム側から読み上げる内容について述べる．列挙する内容は，検索条件（フェーズ I），レストラン名（フェーズ II），レストラン詳細情報（フェーズ III）の 3 つである．これらを**表 2** にまとめる．フェーズ I では，料理の種類・場所・予算のいずれのスロットも埋まっていない場合は，まずこの 3 つのスロット名を列挙し，ユーザに条件を入力したいスロットを指定させる．具体的には図 1 内の例のような内容を列挙する．次に，音声認識部の言語モデルを，その指定したスロットの内容のみを語彙とするものに切り替え，ユーザ発話を認識する．言語モデルを限定してもなおユーザ発話の理解結果が得られない場合は，当該スロットの詳細を列挙する．この際，料理の種類・場所・予算は列挙する内容が多いため，中間カテゴリを人手で設定した．表 2 中の列挙される語の例の，1 コラム目が中間カテゴリで，2 コラム目がカテゴリ内の内容を表している．列挙の際はまず中間カテゴリの内容を列挙し，さらにそのカテゴリ内の内容を列挙する．つまり料理の種類の中間カテゴリに含まれる内容は「和食」「洋食」などであり，「和食」の中の内容は「寿司」「そば」などである.

さらにフェーズ I では，ユーザ発話に対して WFST による文法検証[26] を行っており，この結果に基づきスロット名の列挙が省略できる場合がある．文法検証では，音声認識結果と，システムが持つすべての文法に対してマッチングがとられる．ある文法の発話であるこ

とを示す正の累積重みと，文法からの逸脱を示す負のペナルティが計算され，その和が正である場合，その和が最大である文法が結果として出力される．負の場合にはどの文法にもマッチしないとして結果は出力されない．文法は，料理の種類・場所・予算に関する発話について，システムが理解可能なユーザ発話に対応する文法を，人手で合計 11 個用意した．複数のスロットを含む文法は文法検証では用いていない．この結果，3.2 節の条件が成立し，列挙される内容がスロット名であった場合において，その直前になされたユーザ発話に対する文法検証結果がたとえば料理の種類を指定する文法であった場合には，システムはスロット名の列挙を省略し，列挙型対話を料理の種類の中間カテゴリの列挙から始める[⋆1]．

フェーズ II のレストラン名は，検索条件に合うレストランのうち 10 件を列挙する．該当件数が 10 件以上あった場合は，今回のシステムでは五十音順に最初の 10 件を列挙した．フェーズ III の詳細情報は，システムが保持するお店の詳細情報（お店の PR ポイントなど，全部で 6 項目）を列挙した．

5. 一般の被験者を用いた評価実験

5.1 実 験 条 件

実験条件は，実環境において公開されたヒューマノイドロボットとの音声対話を想定し，以下の状況となるよう設定した．

(1)　接話型マイクを使用せず音声認識率は高くない．

(2)　ユーザはしばしばシステムにとっての未知語を使用する．

これらについて順に詳しく説明する．

まず 1 点目に対応して，本実験では，マイクロフォンをユーザから 30 から 40 センチメートル離れた位置に設置した．マイクロフォンの配置を**図 3** に示す．システム発話を再生するスピーカは，マイクロフォンから約 25 センチメートルの位置に配置した．これは，ヒューマノイドロボットのスピーカとマイクが近接している状況を想定している．この場合，スピーカから出力されるシステム発話とユーザ発話の混合音がマイクロフォンに入力されるため，Semi-Blind ICA 手法に基づく音源分離処理を行い，ユーザ発話のみを得る[8]．これにより，システムによる選択肢読み上げ中にユーザが割り込んで話すこと（バージイン）を可能としており，ユーザは列挙型対話において任意のタイミングで話すことができる．分離されたユーザ発話を入力としてシステムは音声認識を行い，音声認識結果とユーザの発話開始タイミングを得る．

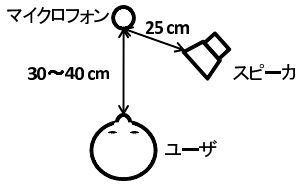

図 3　マイクの配置
Fig. 3　Microphone setting.

表 3　課題の内容
Table 3　Contents of scenarios.

課題	検索条件	詳細情報
1	祇園，鶏料理（和食）	電話番号
2	金閣寺周辺，お好み焼き（和食）	閉店時間
3	京都駅周辺，イタリア料理（洋食）	お店のウリ
4	八坂神社 の近く，予算 3,000 円	電話番号
5	京都大学 の近く，創作料理（和食）	営業時間

この環境での音声認識性能の参考値として，図 3 中のユーザの発話位置からスピーカで JNAS200 文の読み上げデータを再生し，それを録音したデータに対して音声認識率を算出した．元の読み上げデータそのものに対する単語正解率は 92.0%であったのに対し，図 3 の環境で再生し録音した場合の単語正解率は 77.4%であった．この際の音声認識の言語モデル，音響モデルには同一のものを用いている．このように本実験の環境では，接話型マイクを使用しないことから，音声認識性能は高くない．実験時にはシステム発話との音源分離処理による歪みが生じるため，音声認識性能はさらに低下していたと考えられる．

次に 2 点目に対応して，被験者に提示する課題にシステムにとっての未知語を設定した．被験者は**表 3** にある 5 つの課題に沿って対話を行う．具体的には各課題について，検索条件を入力してそれを満たすレストランを 1 つ選択し，そのレストランについて，指定された詳細情報を聞き出す．表中の下線部が，システムにとっての未知語である．**図 4** に示されるように，被験者に未知語が存在することは明示的には知らせていないが，「理解できる言葉や文は限られている」ことは教示した．またシナリオ内で示された地名そのものを，必ず

⋆1 文法検証結果について，評価に用いた全被験者 31 名による対話データのうち 11 名分の対話を抽出して，正誤を人手で集計した．この結果，文法検証の対象となる発話は 63 回あり，そのうち 6 回で文法検証結果が出力されており，5 回が正しく列挙内容を推定できていた．文法検証結果は，出力が誤っていなければ，出力されないことによる損害はないため，1 回の誤りを除けば，ユーザの意図を推定して対話を短縮していた．

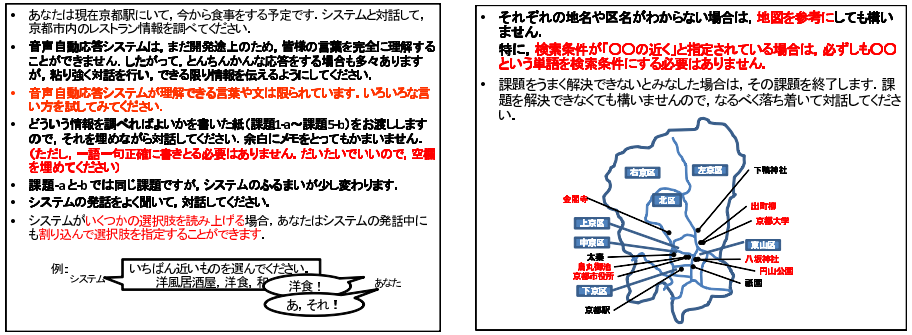

図 4　ユーザに与えた教示（一部）
Fig. 4　Instructions given to participants (excerpt).

しも検索条件とする必要はない旨も伝えた．

ここで被験者に与えた事前教示（図 4）についてさらに説明する．まず被験者への教示は，実験開始前に一括して 1 度だけ行った．具体的には，教示が書かれた紙を渡し，十分よく読んでもらった後に実験を開始した．これは，システムを変えるごとに教示を行うよりも，教示自身や実験者の介入による影響を軽減できると考えたためである．また，被験者は 2 種類のシステムを使用したが，「システムのふるまいが少し変わります」という教示にとどめ，2 つのシステム間の違いが具体的には分からないようにしている．さらに，システムが選択肢を読み上げる場合には，それに対して割り込んで話せることを，簡単な例とともに示した[⋆1]．この例は図 4 左下に示されている．また，被験者には簡単な京都の地図を渡した（図 4 右下）．これは東京近郊で被験者を募集して実験を行ったためで，被験者がある程度京都の地理を知っている場合は不要である．つまり「○○の近く」という検索条件に対して，被験者が代替案をまったく知らないことが予想されたため用意した．

被験者は，最初に簡単な練習セッションを行った後，各課題ごとに，下記の (a)，(b) の 2 つのシステムをこの順に使用し，対話を行った．

(a)　音声認識結果のみに基づきユーザ発話を解釈するシステム．

⋆1 システム (a) でもフェーズ II ではレストラン名が列挙される．ここで被験者が意図的にバージインしながら発話したのは 11 名によるデータ中に 21 件存在した．そのうち 12 件で指示語ではなく内容語が使われており，これらはシステム (a) でも解釈可能である．したがって「割り込んで話せる」という教示により，システム (a) が解釈できない発話が行われたのは 1 名（5 対話）あたり平均 1 回以下であり，この影響は軽微であったと考える．

表 4　課題ごとのタスク達成率（%）被験者 31 名
Table 4　Task success rate per scenario (%) by 31 subjects.

		課題 1	課題 2	課題 3	課題 4	課題 5	合計
全ユーザ	(a) 3 分	19.4	71.0	3.2	22.6	16.1	26.5
全ユーザ	(b) 3 分	35.5	77.4	38.7	45.2	61.3	51.6
	(b) 5 分	58.1	90.3	58.1	74.2	74.2	71.0
列挙型対話を通じた達成	(b) 3 分	20.0	72.0	36.7	26.1	42.9	39.5
	(b) 5 分	45.8	88.0	56.7	65.2	61.9	63.4

(b)　発語行為レベルの情報と音声認識結果を併用してユーザ発話を解釈するシステム．フェーズ I と III では，3.2 節の条件が満たされた場合発語行為レベルの情報を利用する部分対話へと移行する．

つまり (a) での解釈方法は (b) に包含される．このため，実験の途中で，発話タイミングや発話後の沈黙を利用した手法が使用不可能となるのを防ぐため，システムの使用順序は各課題ごとに (a) → (b) の順とした．また対話時間に制限を設け，それぞれ (a) は 3 分，(b) は 5 分とした．(a) を 3 分としたのは，(a) のシステムでは音声認識誤りが繰り返し発生することで対話が停滞することがあり，その場合 5 分間対話を続けさせるのは被験者にとって負担になるためである．また (b) の 3 分の時点での対話の進行状況を参考までに記録した．それぞれの制限時間内で条件を満たすレストランの詳細情報を聞き出せた場合，タスク達成とした．

被験者は 20 代〜60 代の，東京近郊で募集された一般男女 31 名である．被験者から収集した発話の総数は 6,952 発話であり，その平均の単語正解率は 67.8%であった．音声認識には，20 dB のピンクノイズを重畳した音響モデルと，CIAIR コーパス[27] とシステムの保持する検索条件などの単語から作成した統計的言語モデル（語彙サイズ：8,239）を使用した．デコーダには，デコーダ VAD 機能を備えた Julius[28] を使用した．

5.2　実 験 結 果

5.2.1　課題ごとのタスク達成率

発語行為レベルの情報を発話の解釈に使用した場合の影響を調べるために，システム (a) と (b) について各課題ごとのタスク達成率を比較する．まず全 31 ユーザに対する，各課題ごとのタスク達成率を**表 4** に示す．表 4 中の (b) 3 分は，(b) のシステムでの 3 分時点のタスク達成率，(b) 5 分は，5 分時点でのタスク達成率である．音声認識率が高くなく，課題中に未知語が含まれることなどから，全般に時間内のタスク達成率は必ずしも高くない．

ここで，ユーザはシステムを (a) → (b) の順で使用したことから，(b) のタスク達成の要

因にはシステムへの慣れが含まれる．実際，システム (b) のフェーズ I と III で，1 度も列挙型対話に誘導されずに達成された課題が，課題 1 から 5 の順にそれぞれ，7 回（3 分時点では 6 回），6 回，1 回，8 回，10 回存在した．少なくともこれらの場合は，本手法の貢献ではなく慣れの影響によりタスクを達成できたと考えられる．慣れの影響を完全に排除することはできないが，これらの場合を除外し，1 度でも列挙型対話に誘導された場合のタスク達成率を，表 4 の最下段「列挙型対話を通じた達成」の行に示している．この場合を見ると，(b) 3 分によるタスク達成率は，(a) 3 分と同等もしくは向上が見られている．依然，慣れによる影響が完全に除外できているとはいえないが，後述するように，システム (a) では未知語などによる音声認識誤りが繰り返されタスクが達成できなかった場合に，システム (b) で列挙型対話に誘導されることで初めてタスクを達成できた例が多数見られたことをあわせて考えると，提案手法によるタスク達成率の向上は存在したと考えられる．また表 4 最下段の (b) 5 分では，(b) 3 分と比べてタスク達成率が高い．一般には，時間をかければタスク達成率が向上するのは当然であるが，音声認識率が低い場合やユーザ発話に未知語が含まれる場合は，時間をかけても音声認識誤りが繰り返されるだけでタスク達成率の向上につながらない場合がある．実際，システム (a) では，音声認識結果が正しい場合が連続しないことで，3 分の間にタスクが一向に進捗しないケースが見られた[⋆1]．一方，システム (b) では，誤りが続いた場合にはシステム側から選択肢を列挙することからも，時間をかければタスク達成率は向上していたことが分かる．これにより，タスク達成に時間が必要であるものの，発話タイミングや発話後の沈黙を利用する本手法の有効性が示唆されている．

また，被験者が未知語を用いていた場合に (b) で列挙型対話に切り替わることで初めてタスク達成可能となった場合が数多く存在した．実際，(a) ではタスク達成できず，そのうち (b) 3 分で列挙型対話に誘導されることで初めてタスク達成できた対話は 27 対話であった．たとえばある被験者は，課題 2 のフェーズ III において，(a) ではシステムにとっての未知語である語彙 “閉店時間” を尋ね続け，詳細情報を聞き出すことができなかった．一方 (b) では，システムがお店の詳細情報を，「途中で指定してください．お店の PR ポイント，予算，住所，最寄り駅，電話番号，営業時間」のように列挙した．この結果，被験者が自分の意図に対応するシステムの語彙 “営業時間” を使うことで，詳細情報を聞き出せていた．ユーザにとっては，システムの語彙内の単語だが音声認識誤りが起こっているのか，語彙外

⋆1 そのような状態で対話をさらに続けるのは被験者にとって負担が大きいと考え，システム (a) をさらに長く使った場合の実験は行わなかった．

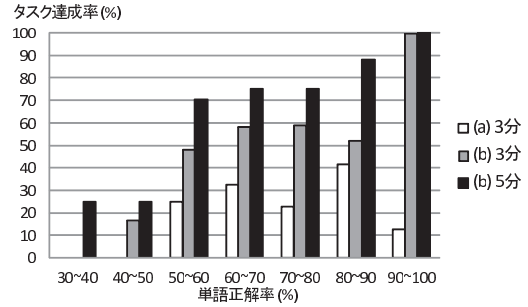

図 5　単語正解率ごとのシステム (a)，(b) のタスク達成率
Fig. 5　Task success rate for two systems per ASR accuracy.

だから正しく認識されないのかは区別できないため，システム側からの語彙の提示やそれに対するタイミングを使った指定が有効であったと考えられる．

また，タイミングの併用による性能向上を検証するために，5.2.3 項で後述する 11 名分のデータにおいて，フェーズ II での，システム (a)，(b) による指示対象同定精度を調べた．つまり (a) では音声認識結果のみで，(b) では発話タイミングを併用して指示対象を同定していた．ユーザが列挙項目中の対象を意図的に指示する発話は合計 57 発話存在し，指示対象同定精度はそれぞれ，システム (a) で 43%，システム (b) で 83%であった．これより列挙型対話での指示対象同定において，発話タイミングを併用する効果が示されている．そもそもユーザが指示語を用いた場合には発話タイミングでしか解釈できないが，以前の実験では，ユーザが内容語を発話した場合でも，発話タイミングの併用による指示対象同定精度の向上が報告されている[23]．

5.2.2　単語正解率とタスク達成率の関係

被験者の単語正解率とタスク達成率との関係を**図 5** に示す．図 5 では，各単語正解率ごとのタスク達成率を，(a)，(b) 3 分，(b) 5 分のそれぞれの条件ごとに示している．この図から，同程度の単語正解率でも，(a) に比べて (b) の両条件では，タスク達成率が高いことが分かる．(a) で単語正解率が高くてもタスク達成できていない場合があるのは，たとえば被験者がレストランの検索条件に “雰囲気のいいところ” という条件を追加しようとしており，その音声認識結果が正しくてもシステムが理解できる範囲を超えた表現であったことが原因だった．つまり，音声認識だけではなく，言語理解レベルでの未知語も，タスク達成率を下げていたことが分かる．このことも，システム側から必要な場合に選択肢を列挙すると

表 5　タスクを達成した被験者の単語正解率（%）の最小値
Table 5　Minimum ASR accuracy by subject who achieved task.

課題	1	2	3	4	5
(a) 3 分	56.1	57.3	65.6	53.1	59.3
(b) 3 分	52.6	46.2	66.1	48.2	54.4
(b) 5 分	36.9	46.2	52.0	46.7	54.4

いうアプローチにより，ユーザとシステムの間の語彙の乖離が埋められていた可能性を示唆している．

さらに，タスクを達成した被験者の中での，単語正解率の最小値を調査した．各課題ごとに分類した結果を**表 5** に示す．表より，(a) 3 分よりも，ほぼ同等である課題 3 以外は (b) 3 分の方が最小値は小さく，さらに (b) 5 分ではすべての課題において最小値は小さい．このことから音声認識率が低くても，システム (b) の場合には時間をかければタスク達成が可能であったことが分かる．

5.2.3　発話後の沈黙を用いた解釈に関する考察

全 31 名の被験者によるデータのうち，11 名分を抽出し，人手でタグをつけてさらに詳しく分析を行った．なおこの 11 名は，システム (b) 5 分でのタスク達成数の平均が全被験者の平均とおおよそ同じになるよう抽出した．

この 11 名分の対話データに対して，発話後の沈黙を用いた解釈の成否について調査した．具体的には，この解釈を行う際のプロンプトがすべて再生された後に，ユーザが肯定の意図を持っていた場合と否定の意図を持っていた場合を人手で判定し，それぞれの回数を計数した．肯定の意図を持っていたと思われる場合に，実際に沈黙し，正しく肯定の意図を伝えていたのは 49 例中 49 例で，ここではすべての場合で成功していた．一方，否定の意図を持っていたと思われる場合に，被験者が実際に何らかの発話を行い，否定（または訂正）の意図を正しく伝えていたのは，26 例中 19 例であった．正しく解釈されなかった 7 例では，被験者が否定または訂正の意図を持って発話したにもかかわらず，処理の遅延などにより，それが 3 秒以内と判定されなかったものである．

これをふまえて，ユーザの沈黙を引き出すプロンプトの表現や制限時間の設定に工夫が必要である．本システムでは実験的に 3 秒の沈黙をもって肯定と見なしたが，この時間はシステムのタスクや被験者の好みに依存する可能性が示唆されている．被験者へのアンケート中に，制限時間を何秒に設定すればよいかという設問を設定したところ，実際「3 秒でちょうどいい」が 6 名，「3 秒より長い方がいい」が 3 名，「3 秒より短い方がいい」が 3 名という結果であった．アンケートへ回答したユーザについて，対話中のシステム発話へのバージイン回数と回答結果の関係を調査した．「3 秒より長い方がいい」と回答した被験者は，1 課題中平均で 20.3 回発話があり，そのうち平均 2.7 回バージインをしていた．一方，「3 秒より短い方がいい」と回答した被験者は，1 課題中平均で 21.1 発話があり，そのうち平均 5.0 回バージインをしていた．つまり，バージインしない被験者ほど長い制限時間を好むという傾向がみられた．これは，バージインしない被験者ほどシステム発話をよく聞いてから応答する傾向があり，制限時間が短いと応答内容を考えている間にシステムが次の発話に移ってしまうという場合があったと考えられる．つまり，発話後の沈黙を用いた解釈にユーザ適応の余地があり，そのための特徴として当該ユーザのバージインの頻度が有用である可能性が示唆された．

最後に，発話タイミングを用いた解釈と発話後の沈黙を用いた解釈による，タスク達成率の改善に対する寄与の大小について考察する．抽出した 11 名とシステム (b) との対話において，フェーズ I および III で列挙型対話における指示対象同定が成功した回数は 94 回，発話後の沈黙を用いた解釈が成功した回数は 68 回であった．これより前者による貢献の方が相対的に大きかったことが類推できる．定性的には，列挙型対話における発話タイミングを用いた解釈は，未知語など，ユーザとシステム間に語彙のギャップが存在する局面で特に有効であったと考えられる．一方発話後の沈黙を用いた解釈は，音声対話システムでは確認に対する肯定・否定の認識を誤ると対話の破綻につながるため，特に音声認識が困難な状況において有効であったと考えられる．

6. ま と め

本研究では，発語行為レベルの情報をユーザ発話の解釈に用いた音声対話システムを構築し，一般から募集した被験者 31 名による対話実験を行った．実験の結果から，音声認識率が低い場合や，ユーザ発話に未知語が含まれる場合でも，提案手法によるタスク達成率の向上が示唆された．本稿での実験設定は，一般的な音声対話システムの研究の評価実験とは異なり，接話型マイクを用いないなど，ヒューマノイドロボットとの実環境下での対話を強く意識したものである．このため，やや特殊な条件下で対話実験を行ったが，これは実ユーザに対してシステムを公開する際には十分にありうる条件であると考える．5 章の実験では，システムの使用順を (a) → (b) と固定したうえで，事後的にその慣れによる影響の排除を試みた．しかし慣れの影響を十分に取り除き，実験結果の信頼性を増すには，使用順を被験者群ごとに入れ替えるといった実験計画に基づく，さらなる検証が必要である．

2 章で述べたように，本研究は自然なターンテイキングの実現を目指したものではなく，音声認識が困難な状況下で，発話のタイミングや発話後の沈黙を使うことで，対話の成立を目指した試みである．特に，発話後の沈黙を用いた解釈では，ユーザは意図的にシステムのタイムアウトを待つことから，人間同士が行うような自然な対話ではない．しかし一方で，被験者へのアンケート結果によると，31 名中 28 名が実際にこの部分対話を経験し，そのうちの 23 名はこの機能に対して 5 段階中 4 以上の肯定的な評価を与えた．付随する自由記述では，「システムが認識できないのであれば，早めに教えてくれるのはいいと思う」「対話が進むのであったほうがよいと思う」などの感想が得られた．このように，音声認識結果のみに基づく解釈だけでは立ち行かない場合に，音声に付随する他の情報を活用して，ユーザ発話を解釈することの有用性が示されている．

今後の課題として，まず列挙型対話における列挙内容の動的な選択があげられる．列挙内容の選択は対象ドメインに強く依存する部分であり，本稿のシステムでは基本的に事前に人手で設定した順番および内容で選択肢を列挙した．この列挙内容の順序は，ユーザや状況に応じて柔軟に決定できるのが望ましい．また表 2 では，地名など列挙すべき内容が多い場合は，中間カテゴリを設定していったん絞り込むようにした．しかし適切なカテゴリが設定できない場合は，1 度に列挙する項目数が多くなり，列挙にさらに時間がかかるという問題がある．項目数が多い場合はすべてを逐一列挙することはできないため，この点は選択肢を列挙するというアプローチの限界である．

また発話タイミングの検出は，音声認識エンジンの発話区間検出結果を用いている．本稿では，発話区間検出性能の高い，デコーダ VAD 機能を備えた Julius[28] を用いたが，発話区間検出性能による解釈結果への影響の評価や，それに応じた手法の拡張も検討が必要である．

さらに列挙型対話への切替え条件の最適化があげられる．本稿では，表 1 にあげたように，雑音比や対話履歴の特徴に対して経験的にしきい値を設定し，列挙型対話への切替え条件とした．V_{SNR} は音響的な特徴であるためドメインには依存しないが，3.2 節や 3.3 節で述べた繰返しの回数やタイミングの設定は，ユーザの特性や状況に応じて異なる可能性がある．抽出した 11 名分のデータに対して人手による予備的な評価を行った結果，フェーズ I において列挙型対話へ移行したケースはおおよそ妥当であったものの，フェーズ III では $C_{listup} > 2$ の条件により，過剰に列挙型対話へ移行している様子も観察された．得られたデータに基づく機械学習による移行判定の最適化も，今後の課題としてあげられる．

謝辞　京都大学学術情報メディアセンターの森准教授，同情報学研究科の笹田氏には，言語モデルの作成に関してご助力いただいた．ここに記して感謝する．本研究の一部は科研費の支援を受けた．

参 考 文 献

1) 竹林洋一：音声自由対話システム TOSBURG II—ユーザ中心のマルチモーダルインタフェースの実現に向けて，電子情報通信学会論文誌，Vol.J77-D-II, No.8, pp.1417–1428 (1994).
2) Komatani, K., Ueno, S., Kawahara, T. and Okuno, H.G.: User Modeling in Spoken Dialogue Systems to Generate Flexible Guidance, *User Modeling and User-Adapted Interaction*, Vol.15, No.1, pp.169–183 (2005).
3) Raux, A., Bohus, D., Langner, B., Black, A.W. and Eskenazi, M.: Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience, *Proc. Int'l Conf. Spoken Language Processing* (*INTERSPEECH*) (2006).
4) 翠　輝久，河原達也，正司哲朗，美濃導彦：質問応答・情報推薦機能を備えた音声による情報案内システム，情報処理学会論文誌，Vol.48, No.12, pp.3602–3611 (2007).
5) 鹿野清宏，Tobias, C.，川波弘道，西村竜一，李　晃伸：音声情報案内システム「たけまるくん」および「キタちゃん」の開発，情報処理学会研究報告，2006-SLP-63-7 (2007).
6) Austin, J.: *How to Do Things with Words*, Oxford University Press (1962).
7) Searle, J.R.: *Speech Acts*, Cambridge University Press (1969).
8) 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 International Conference on Intelligent Robots and Systems* (*IROS*), pp.1718–1723 (2008).
9) 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，2009-SLP-76-14 (2009).
10) 横山真男，青山一美，菊池英明，白井克彦：人間とロボットのコミュニケーションにおける非言語情報の利用，情報処理学会研究報告，98-SLP-21-7, pp.69–74 (1998).
11) 藤原敬記，伊藤敏彦，荒木健治：タスク指向対話における発話意図の対話リズムへの影響，情報処理学会研究報告，2007-SLP-66-7, pp.37–42 (2007).
12) 豊倉正佳，翠　輝久，河原達也：音声対話システムにおける対話相手・発話意図と発話タイミングの関係の分析，人工知能学会研究会資料，SIG-SLUD-A702-4, pp.21–28 (2007).
13) Ström, N. and Seneff, S.: Intelligent barge-in in conversational systems, *Proc. Int'l Conf. Spoken Language Processing* (*ICSLP*), pp.652–655 (2000).
14) Rose, R.C. and Kim, H.K.: A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems, *Proc. IEEE Automatic Speech Recognition*

and Understanding Workshop (*ASRU*), pp.198–203 (2003).
15) Ljolje, A. and Goffin, V.: Discriminative Training of Multi-Stage Barge-in Models, *Proc. IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU*), pp.353–358 (2007).
16) 岡登洋平，加藤佳司，山本幹雄，板橋秀一：韻律情報を用いた相槌の挿入，情報処理学会論文誌，Vol.40, No.2, pp.469–478 (1999).
17) Nakano, M., Dohsaka, K., Miyazaki, N., Hirasawa, J., Tamoto, M., Kawamori, M., Sugiyama, A. and Kawabata, T.: Handling Rich Turn-Taking in Spoken Dialogue Systems, *Proc. European Conf. Speech Commun. & Tech.* (*EUROSPEECH*), pp.1167–1170 (1999).
18) Kitaoka, N., Takeuchi, M., Nishimura, R. and Nakagawa, S.: Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems, *Journal of The Japanese Society for Artificial Intelligence*, Vol.20, No.3, pp.220–228 (2005).
19) 藤江真也，福島健太，三宅梨帆，小林哲則：相槌生成/認識機能を持つ音声対話システム，人工知能学会研究会資料，SIG-SLUD-A502-09, pp.41–46 (2006).
20) Raux, A. and Eskenazi, M.: A Finite-State Turn-Taking Model for Spoken Dialog Systems, *Proc. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT NAACL*), pp.629–637 (2009).
21) 平沢純一，宮崎　昇，相川清明：質問–応答連鎖からの音声対話システムの誤解の検出，情報処理学会研究報告，2000-SLP-34-41, pp.239–244 (2000).
22) 田中一晶，左　祥，嵯峨野泰明，荒木雅弘，岡　夏樹：No News 規準が有効な条件：誘導教示の意味学習場面での実験的検討，電子情報通信学会論文誌，Vol.J92-A, No.11, pp.784–794 (2009).
23) 松山匡子，駒谷和範，武田　龍，尾形哲也，奥乃　博：バージイン許容音声対話システムにおけるユーザ発話の分析と指示対象同定への応用，情報処理学会研究報告，2010-SLP-82-21 (2010).
24) Kim, C. and Stern, R.M.: Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis, *Proc. Annual Conference of the International Speech Communication Association* (*INTERSPEECH*), pp.2598–2601 (2008).
25) 神田直之，駒谷和範，尾形哲也，奥乃　博：データベース検索タスクにおける対話文脈を利用した音声 言語理解，情報処理学会論文誌，Vol.47, No.6, pp.1802–1811 (2006).
26) Komatani, K., Fukubayashi, Y., Ikeda, S., Ogata, T. and Okuno, H.G.: Selecting Help Messages by using Robust Grammar Verification for Handling Out-of-Grammar Utterances in Spoken Dialogue Systems, *IEICE Trans. Information and Systems*, Vol.E93-D, No.12, pp.3359–3367 (2010).
27) 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，情報処理学会研究報告，2003-SLP-49-24 (2003).
28) 李　晃伸：大語彙連続音声認識エンジン Julius ver.4，情報処理学会研究報告，2007-SLP-69-53, pp.307–312 (2007).

(平成 23 年 4 月 11 日受付)
(平成 23 年 9 月 12 日採録)

駒谷　和範（正会員）

1998 年京都大学工学部情報工学科卒業．2000 年同大学院情報学研究科知能情報学専攻修士課程修了．2002 年同大学院博士後期課程修了．京都大学博士（情報学）．同年京都大学大学院情報学研究科助手．2007 年同助教．2010 年より名古屋大学大学院工学研究科准教授．同年より JST さきがけ「情報環境と人」領域研究員兼務．主に音声対話システムの研究に従事．2008 年から 2009 年まで米国カーネギーメロン大学客員研究員．情報処理学会平成 16 年度山下記念研究賞，FIT2002 ヤングリサーチャー賞等を受賞．電子情報通信学会，言語処理学会，人工知能学会，ISCA 各会員．

松山　匡子

2009 年京都大学工学部情報学科卒業．2011 年同大学院情報学研究科知能情報学専攻修士課程修了．現在パナソニック株式会社勤務．在学中はバージインタイミングを活用した音声対話システムの研究に従事．情報処理学会平成 21 年度音声言語処理研究会（SIG-SLP）学生奨励賞，2010 年度人工知能学会研究会優秀賞を受賞．

武田　　龍

2006 年京都大学工学部情報学科卒業．2008 年同大学院情報学研究科知能情報学専攻修士課程修了．2011 年同大学院博士後期課程修了．京都大学博士（情報学）．現在，株式会社日立製作所勤務．在学中は複数話者に対する音声認識や音源分離の研究に従事．IEEE/RSJ IROS2006 Best Paper Nomination Finalist.

高橋　　徹（正会員）

1996 年名古屋工業大学知能情報システム学科卒業．2004 年名古屋工業大学大学院工学研究科電気情報工学専攻，博士後期課程修了．博士（工学）．和歌山大学システム工学部産学官連携研究員を経て，2008 年より京都大学大学院情報学研究科グローバル COE 助教．研究分野は，ロボット聴覚および音声コミュニケーション．音声による人間–ロボット間インタラクションのための音声認識・合成．ロボット聴覚ソフトウェア HARK，音声分析変換合成システム STRAIGHT の開発．RSJ，IEICE，ASJ 各会員．

尾形　哲也（正会員）

1993 年早稲田大学理工学部機械工学科卒業．日本学術振興会特別研究員，早稲田大学理工学部助手，理化学研究所脳科学総合研究センター研究員，京都大学大学院情報学研究科講師を経て，2005 年より同助教授（現准教授）．博士（工学）．JST さきがけ研究「情報環境と人」領域研究員（5 年）．この間，早稲田大学ヒューマノイド研究所客員准教授，同大学理工学研究所客員准教授，理化学研究所脳科学総合研究センター客員研究員等を兼務．研究分野は人工神経回路モデルおよび人間とロボットのコミュニケーション発達を考えるインタラクション創発システム情報学．日本ロボット学会，日本機械学会，人工知能学会，計測自動制御学会，ヒューマンインタフェース学会，バイオメカニズム学会，IEEE 等各会員．

奥乃　　博（正会員）

1972 年東京大学教養学部基礎科学科卒業．日本電信電話公社，NTT，JST，東京理科大学を経て，2001 年より京都大学大学院情報学研究科知能情報学専攻教授．博士（工学）．この間，スタンフォード大学客員研究員，東京大学工学部客員助教授．人工知能，音環境理解，ロボット聴覚，音楽情報処理の研究に従事．1990 年度人工知能学会論文賞，IEA/AIE-2001，2005，2010 最優秀論文賞，IEEE/RSJ IROS-2001，2006 Best Paper Nomination Finalist，IROS-2010 NTF Award for Entertainment Robots and Systems，第 2 回船井情報科学振興賞等受賞．本学会理事，人工知能学会，日本ロボット学会，日本ソフトウエア科学会，ACM，IEEE，AAAI，ASA 等各会員．