社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B303-05

実世界知識を扱う音声対話技術とクラウドロボティクスへの展開

Grounded Spoken Dialogues with Robots: Cloud Robotics Tools and Service Robot Applications

杉浦孔明

Komei Sugiura

情報通信研究機構

National Institute of Information and Communications Technology

komei.sugiura@nict.go.jp

Abstract

ロボットとの音声によるコミュニケーションは，ユーザにとって手軽であるというメリットがあるが，実現は簡単ではない．頑健な音声認識が必要とされるだけでなく，発話の解釈が実世界情報や履歴により影響を受けるためである．このような背景から，音声・画像・動作・コンテキスト情報を用いてユーザの発話を解釈するロボット対話技術 LCore を開発している．本稿では，音声対話を通じた実世界知識の学習と行動生成について述べたのち，サービスロボットの競技会であるロボカップ@ホームにおけるロボット対話技術の応用について紹介する．また，ロボットの音声対話機能の技術動向，特にクラウドロボティクスへの展開と課題について述べる．

1 はじめに

人口構造の変化や労働形態の多様化とともに，生活環境で人間と共存・支援するロボットへの期待が高まっている．実際に，掃除ロボットの国内市場規模は 2006 年に約 2 万台であったものが，2012 年には約 38 万台になっている．また，2014 年にはリモートプレゼンスロボットがノート PC 程度の価格帯で発売されたことから，普及が期待される．家庭用ロボットで使われる技術はますます高度化，多目的化すると考えられる．Google 会長のエリック・シュミットらは，当面は高度な機能を有する多機能ロボットは一般消費者には高くて手が届かないものの，将来的に一般家庭でも数台の多目的ロボットを持つようになると予想している [Cohen 13].

一方，どれほど豊富な機能を持つロボットを構築したとしても，人が手軽に機能を使えなければ普及しないであろう．つまり，多機能なロボットが日常生活に浸透するためには，人とのコミュニケーション機能が課題となる [西田 03]．優れた音声コミュニケーション機能を有するロボットが現れれば，日常生活が大きく変わる可能性がある．

ロボットとの音声によるコミュニケーションは，ユーザにとって手軽であるというメリットがあるが，実現は簡単ではない．もちろん，雑音抑圧，発話検出，話者分離などは音声処理における重要な課題である．しかし，本稿で強調したい本質的な課題は，発話の解釈が実世界情報や経験により影響を受けることである．例えば「コップ取って」という命令を実行するには，どのコップを取るのか，「取る」とはどのような動作軌道を表すのか，を推論しなければならない．一方，家族にコップを取ってもらう場合は省略した言い方でも通じることが多いうえ，わからなければ聞き返すだろうという期待がある．つまり，人間はそれまでの経験から，相手が自分の発言をどれくらい理解できるかに関して暗黙的な知識がある．しかし，ロボットと対話するユーザにとって，ロボットが状況をどれだけ理解しているかを推定するのは難しい．

現状のロボット対話処理機構では，動作コマンドの伝達を目的とすることが多いにも関わらず，動作情報と音声認識は別々に処理されることが多い [Hartanto 11]．それに対し，ロボットの対話機構にグラウンドしない知識を導入することが，シンボルグラウンディング問題を孕むことは古くから指摘されている [Harnad 90,Pfeifer 99]．したがって，ロボットに人間と自然にコミュニケーションさせるためには，ユーザや状況への適応手法が重要課題となる．

本稿では，実世界知識を扱う音声対話技術について概説する．まず，ロボットとの音声対話における研究を分類し課題を整理する．3 節では，著者らの研究グループがこれまでに行なってきたロボット対話技術について述べる．4 節では，それらの技術の実世界への応用例として，ロボカップ@ホームタスクへの適用例を紹介する．5 節では，著者らが開発したサービスロボット向けのクラウド型音声コミュニケーションツールキットについて説明する．

2 ロボットとの音声対話

実世界にグラウンドした対話を実現するためには，実世界のオブジェクトや動作を記号化・言語化することが極めて重要な課題である．ロボティクスの分野では，動作-記号の相互変換に関する試みが近年注目されてきている [Inamura 04, Ogata 07, 高野 09]．高野らは，運動の分節化を通じてヒューマノイドロボットが獲得した原始シンボルを用いて，運動認識・生成を行っている [高野 09]．Ogata らは，動作系列と記号列の間の多対多対応問題を扱うリカレントニューラルネットに基づく手法を提案している [Ogata 07]．Kollar らは，ロボットに与える移動指示に関して，ランドマークオブジェクトや動作にグラウンドした言語表現を学習する手法を提案した [Kollar 10]．学習されたモデルを用いることにより，指示が示す最も確からしい経路を推論する．

オブジェクト-記号の相互変換に関しては，人工知能分野で多くの研究が行われてきた [山肩 04, Roy 02, Dale 95]．詳細については，[長井 12] が詳しい．山肩らはコップ類の名称とイメージモデルの参照関係における曖昧性に一貫した個人差があることを示している [山肩 04]．Roy は，ディスプレイ上の複数の長方形のうちひとつを指示する言語表現をテキストベースで生成する手法を提案している [Roy 02]．[Roy 02] で提案された手法では，単語のカテゴリは教師なし学習の枠組みでクラスタリングされるため，設計者が属性を用意する必要がない．Yu らは，動画と文を入力として，文節が動画中のどの領域に対応するかを学習させている [Yu 13].

音声対話を行うロボットの先駆的事例としては，Jijo-2 が挙げられる [松井 00]．また，稲邑らは，移動ロボットの障害物回避において，ロボットが段階的に行動決定モデルを獲得する機構を提案した [稲邑 01]．センサ値と行動の関係をベイジアンネットを用いてモデル化し，推論結果の確信度を用いて応答決定を行う．

一方，対話システムの分野では，ホテル検索やバスの経路検索などが代表的なタスクである [Komatani 00, Bohus 06, Kawahara 98]．対話システムの評価タスクとしては，1990 年に始まった Loebner Prize[1]がある．Loebner Prize では，端末を通じてシステムと人間が対話を行い，チューリングテストに合格した場合は 10 万ドルが与えられる．これまでに合格したシステムはないものの，毎年最も人間に近い動作をしたと判定されたシステムには賞金が与えられる．音声対話システムの評価タスクとしては，2010 年に行われた “Spoken Dialog Challenge” [Black 11] がある．システムのタスクはバスの経路案内であり，実際に自動応答サービスとして実装された．システムの評価尺度として，単語誤り率やタスク達成率などが用いられている．Spoken Dialog Challenge の後継として，REAL Challenge[2]が企画されている．

スマートフォンを始めとする種々のデバイスに音声インタフェースが導入され，広く一般に認知されるようになってきた [河原 13, 松田 13]．検索や対話に代表されるサービスの多くは，クラウド型サービスとして実装されている．ロボティクスにおいて，クラウド型サービスの利活用を目指す分野はクラウドロボティクス [Kuffner 10] と呼ばれる．代表的な研究としては，Google Goggles を用いたマニピュレーション [Kehoe 13] や，クラウド型知識共有を行うプラットフォームおよびインタフェース言語 RoboEarth [Tenorth 12] などがある．また，著者らはクラウド型の音声コミュニケーションツールキット rospeex を公開している [杉浦 13b].

3 実世界にグラウンドした音声対話

ロボットが生活環境に浸透するに従い，ロボットのコミュニケーション能力が課題となる．現状の技術では,「コップ（テーブルに）置いて」などの曖昧な発話の解釈は非常に難しい．省略された語が表すオブジェクトを推定する必要があるうえ，環境中に存在する「コップ」の候補から正しいオブジェクトを選択しなければならない．

本節では，著者らが開発してきたコミュニケーション学習基盤 LCore [Iwahashi 09] を紹介する．LCore は，画像，動作，アフォーダンス，履歴などの実世界情報を学習し，発話・動作の生成が可能である．

3.1 コミュニケーション学習基盤 LCore

現状のロボットの対話処理機構では，動作コマンドの伝達を目的とすることが多いにも関わらず，動作情報と音声認識は別々に処理されていることが多い [Hartanto 11]．ユーザの発話の意味はグラウンドされない知識に基づいて解釈されるため，動作が状況にふさわしいかどうかは音声認識時には考慮されない．しかしながらこのような手法では，ユーザの発話の意味が状況に応じて適切に理解されない，という問題がある．

LCore では，マルチモーダル入力（音声・画像・予測動作など）から学習されたモデルを用いてユーザの発話を理解する．以下では，各モダリティに対応するモデルをモジュールと呼ぶこととする．

まず，音声発話 s から最適行動 $\hat{a}$ を出力する場合について考える．マルチモーダル発話理解スコアを表す関数 Ψ を，各モジュールの重み付き和として定義する．

$$\Psi(s,a_k,O,q^{\langle i_t \rangle}) = \max_z (\gamma_1 B_S + \gamma_2 B_I + \gamma_3 B_M + \gamma_4 B_R + \gamma_5 B_H) \quad (1)$$

[1] http://www.loebner.net/Prizef/loebner-prize.html

[2] https://dialrc.org/realchallenge/

ここに，z は各単語の概念構造への分割であり，$(s, a_k, O, q^{\langle i_t \rangle})$ はそれぞれ，発話，行動，状況，行動コンテキストを表す．また，$\gamma = (\gamma_1, ..., \gamma_5)$ は各モジュールに対する重みであり，MCE 学習 [Katagiri 98] を用いて学習される．各モジュールが出力するスコアは以下のように定義される．

- 音声スコア B_S
  単語の n-gram，および節の連接確率として学習される．B_S は，発話 s に対する概念構造 z の条件付き確率の対数として表す．
- 視覚スコア B_I
  ガウス分布により学習される．B_I は，オブジェクトの視覚特徴量が与えられたときの対数尤度である．
- 予測動作スコア B_M
  Reference-Point-Dependent HMM（RPD-HMM）[Sugiura 11a] により学習される．B_M は，可能な行動に対して軌道を仮想的に生成したうえで，その軌道の尤度として得られる．
- 動作-オブジェクト関係スコア B_R
  「平らなものはオブジェクトを載せられやすい」など，動作と視覚的特徴の関係を表す．B_R は，2 個のオブジェクトの視覚特徴量に対するガウス分布の対数尤度である．
- 行動コンテキストスコア B_H
  B_H は，あるコンテキスト（「把持されている」,「直前に操作された」など）のもとでの指示対象としてのオブジェクトの適切さ (スコア) を表す．

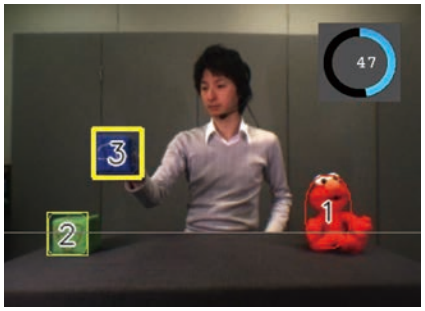
以上より，コンテキスト q，状況 O，発話 s が与えられたときの最適行動 $\hat{a}$ は以下で得られる．

$$\hat{a} = \underset{k}{\operatorname{argmax}} \Psi(s, a_k, O, q) \tag{2}$$

3.2 確信度に基づく動作と発話の生成

前節までの手法は，ユーザから発話が入力され，ロボットが動作を出力するという一方向的過程であった．本節では，入力された発話に対し，ロボットが応答（発話または行動）を出力する場合について考える．例えば，ユーザが「コップ（テーブルに）置いて」などの曖昧な発話を行ったとする．曖昧性を解消するためには,「赤いコップですか，青いコップですか」のような確認発話を毎回行なってもよいが，曖昧なときのみ確認発話を行う方が望ましい．

著者らは，発話理解確率（発話を正しく解釈できる確率）を推定し，効用最大化により応答を生成する手法を提案した [Sugiura 11b]．発話が曖昧であるとき，共有信念関数の第 1 候補と第 2 候補のスコアの差（マージン）が小さいことから，これを曖昧性の尺度として用いている．



【状況】オブジェクト 2 が直前に操作された
U: ハコ エルモ ちかづけて.
R: ミドリハコをちかづけて？
U: いいえ.
R: アオイハコをちかづけて？
U: はい.
R: (動作実行：オブジェクト 3 をオブジェクト 1 に近づける)

図 1: グラウンドした語彙による確認発話の生成

提案手法では，統合確信度を発話理解確率の推定値としてモデル化した．統合確信度は，ベイズロジスティック回帰により学習される．動作応答 b_1 と確認発話応答 b_2 は，対応する期待効用 $\mathbb{E}[R_i](i = 1, 2)$ の最大化により選択される．

$$\mathbb{E}[R_i] = r_{i1} f(d; w) + r_{i2}(1 - f(d; w)) \tag{3}$$

$$d = \Psi(s, \hat{a}, O, q) - \max_{j \neq k} \Psi(s, a_j, O, q) \tag{4}$$

ここに，$f(d; w)$ は，$w = (w_0, w_1)$ をパラメータとするロジスティックシグモイド関数である．また，r_{i1}, r_{i2} はそれぞれ，行動 $\hat{a}$ がそれぞれ正解，不正解のときの応答 b_i に対する効用である．

実験では，オブジェクトを操作するよう，ユーザはロボットに指示を与える．両者は音声対話により曖昧性を解消し，ロボットがユーザの意図した行動をとればタスク成功とした．ロボットが失敗行動（正解でない行動）を行ったときは，アームに取り付けたセンサを叩くことで教師信号を与えることができる．このようにして得たマージンと教師信号の組を用いて，ベイズロジスティック回帰により $f(d)$ を学習させた．

図 1 は，ユーザ (U) とロボット (R) の対話例を示したものである．図において，右上の数値は統合確信度 $f(d)$ を表す．図 1 では，最適行動の確信度は $f(d) = 0.478$ であり，確認発話「アオイ ハコをちかづけて」が最適応答であった．この言語表現は，オブジェクト 2 と 3 の視覚的特徴のなかで最も異なる属性[3]について述べており，ユーザにとって理解しやすい．ランドマークについては確認発話を行わなくても確信度に影響はないため，確認を省略していると考えられる．

[3] カラー画像ではオブジェクト 2 は緑，オブジェクト 3 は青色である．

図 2: 2012 年世界大会に参加したロボット

実験の結果，動作や視覚などの情報を用いず，音声のみで発話理解を行った場合の行動失敗率は 83.4%であった．ベースライン手法（発話を行わず動作のみで応答する）における行動失敗率は 12.0%である一方，提案手法による行動失敗率は 2.6%であった．このことから，提案手法はベースライン手法に比べて行動失敗率を大幅に低減できたといえる．

4 実世界への適用：ロボカップ@ホーム

サービスロボットの研究開発においては，独自の環境や評価尺度が用いられることが多く，一般的に手法同士の比較が難しい．一方，タスクを標準化することで比較評価のコストを低減すれば，コミュニティ全体の研究開発に貢献できるであろう．ロボカップ@ホームは，サッカーやレスキューと並ぶロボカップ [浅田 10] のリーグのひとつであり，生活支援ロボットの競技である [Iocchi 10, 杉浦 12]．各チームのロボットは，日用品の探索，棚からユーザに言われたものを取ってくる，人を追従する等，日常生活に役立つ機能を制限時間内にどれだけ達成できるかを競う．

4.1 ロボカップ@ホームとは

2012 年のロボカップ@ホーム世界大会に出場したロボットを図 2 に示す．ロボカップ@ホームでは，家庭・オフィス・スーパーマーケットなどにおけるロボットの応用を想定したタスクが設定されている．中心課題は，モバイルマニピュレーションとヒューマンロボットインタラクション (HRI) である．後述するように，未知環境における地図作成・移動，日用品の物体認識・把持，高騒音環境における音声認識などを含む．各タスクはベンチマークテストとして明文化されると同時に，複数の技術的課題を含んだストーリーになっており，観客を飽きさせないよう努力されている．

ロボカップ@ホームは 2006 年に始まり，我々は 2008 年より参加してきた．2009 年からは，世界大会のルール策定やジャパンオープン (日本大会) の運営にも参与している．2012 年メキシコで開催された世界大会には 9 カ国から 18 チームの参加があり，ジャパンオープンには 3 カ国から 10 チームの参加があった．各チームは 6〜10 人程度で構成されていることが多い．

世界大会では，2 日間のセットアップ期間にフィールドやオブジェクトが発表されるので，参加者は環境地図構築，オブジェクト登録を事前に行うことができる．マニピュレーションの対象であるオブジェクトは，ペットボトルや菓子などの日用品である．各オブジェクトには，名称（「コーンフレーク」など）とカテゴリ名（「食べ物」など）が定義されている．

ロボカップ@ホームに関する日本語による文献には [岡田 10, 杉浦 12] などがある．また，2011 年世界大会については [Stückler 12] が詳しい．これまでの世界大会における得点傾向が [Iocchi 10] にまとめられているほか，各チームの獲得スコア情報が大会ウェブサイト上で公開されている．また，インターネット上にアップロードされた過去の大会の動画は，イメージをつかむ手段として効果的である．最新版の公式ルールは公式サイト[4]からダウンロードできる．

4.2 タスク環境

タスク環境として 2LDK 程度のモデルルームが用意され，部屋構成や家具・食器等は毎年変更される．2012 年世界大会 (メキシコ) で用いられたタスク環境を図 3 に示す．図 3(a)(b) に示す環境は，9 種類のタスクのうち 7 種類を行うメイン環境（以下,「フィールド」と呼ぶ）である．2012 年世界大会のフィールドは，ロビー，リビングルーム，キッチン，ベッドルームの 4 部屋から構成されている．

実際の使用シーンを想定した環境で性能評価を行う意図から，一部のタスクは店舗などフィールド外で行われる．2012 年世界大会では，Restaurant タスクをレストラン (図 3(c)) で行なった．また，2010 年は玩具店，2011 年はスーパーマーケットにおいてタスクを行なっている．これらのタスクの主眼は，未知環境でのオンライン SLAM 機能とモバイルマニピュレーション機能の評価であるため，事前に環境地図を作成することは許可されていない．図 3(c) の環境にはガラスの仕切りや金属製のチェアなどが存在するため，測距センサのレーザが透過あるいは反射してしまい，環境地図構築が非常に難しい．Follow Me タスク (図 3(d)) では，100 人以上の観客が「ノイズ」となり，音声認識，顔認識，人追従を難しくしている．

4.3 日用品マニピュレーションの模倣学習

家庭内でタスクを行うロボットにとっては,「食器棚からコップを取り出す」などの物体の操作は必要不可欠な機能であり，これらの動作を言語で指示できることが望ましい．一方，各種の日用品や棚に対応する動作を事前にプログラムするコストは非常に大きいうえ，事前にプロ

[4] http://www.robocupathome.org/rules

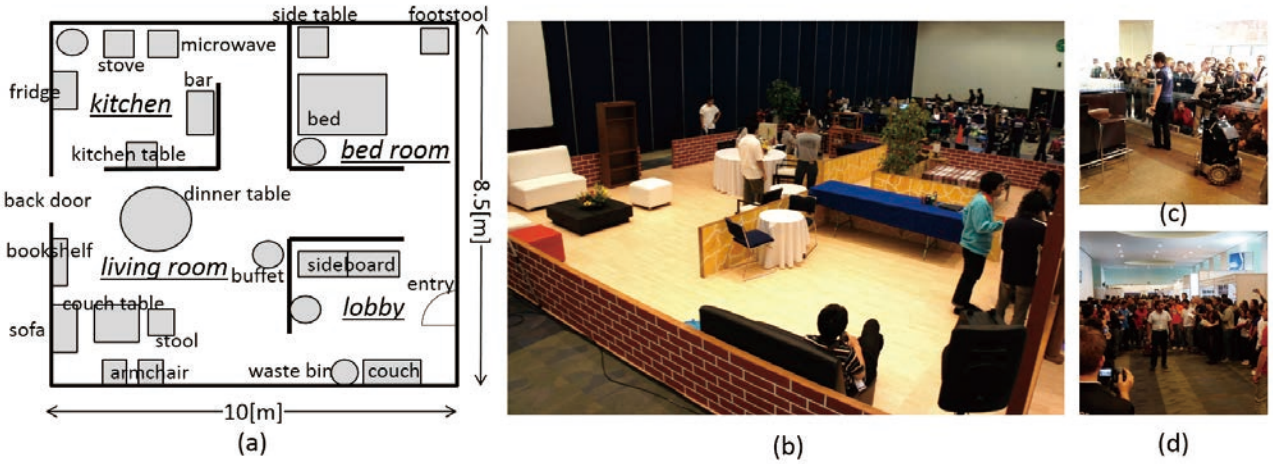


図 3: 2012 年世界大会のタスク環境．(a) 家具配置，(b) メイン環境（フィールド），(c) Restaurant タスクの環境，(d) Follow Me タスクの環境.

図 4: ロボカップ@ホーム環境における動作「捨てる」の学習

グラムされた動作がユーザにとってイメージしにくいものであった場合，安心して動作指示できないという問題がある．そこで我々は，模倣学習の枠組みにより物体操作を学習する手法の開発を行なってきた．このような学習手法を構築することで，プログラミングスキルが必要とされないユーザフレンドリな動作教示方法を実現できる．ロボカップ@ホーム環境における家事動作の学習への適用例を図 4 に示す.

「X を Y にのせる」や「Z を回す」など参照点に依存した動作の模倣では，世界座標系での動作軌道の模倣に意味はなく，適切な座標系を推定し軌道を汎化しなければならない．このために，参照点に依存した隠れマルコフモデル（RPD-HMM）を開発した [Sugiura 11a]．RPD-HMM は，物体位置の時系列を入力として，動作をモデル化するための最適な座標系を EM アルゴリズムにより推定し，軌道のモデルを学習している.

動作の生成時には，学習時とまったく同じように物体が配置されている訳ではないため，たとえ同じ「載せる」動作であっても，学習時の軌道をそのまま用いることは無意味である．学習した RPD-HMM は固有座標系において汎化されたものであるので，固有座標系 C から世界座標系 W へ変換する．RPD-HMM の位置・速度・加速度の平均ベクトルおよび共分散行列は，同時変換行列により C から W 上のモデルに変換される．HMM から滑らかな軌道を生成するために，音声合成の分野で用いられる HMM 軌道生成 [Tokuda 00] を用いる．実験の結果，7 回程度の教示で生成軌道の誤差が収束することが確認できた.

5 クラウド型音声コミュニケーションツールキット “rospeex”

近年，音声対話システムの分野では，開発者が容易に利用できるツールキットが公開されている（例えば [大浦 13]）が，人とロボットのインタラクションでは，高性能な音声認識・合成を容易に利用できる状況ではない．ロボットとの高度な音声インタラクションを可能とするためには，音声処理とロボティクスの深い知識を要求されるのが現状である．このような背景のもと，著者らは音声対話機能の開発コストを下げるべく，クラウド型音声対話ツールキット “rospeex” を開発・公開している．ロボカップ@ホームなどのサービスロボット開発では，開発コストを低減できることから，RT ミドルウェアや ROS（Robot Operating System）などのミドルウェアの利用が一般的になってきている．rospeex は ROS 上で利用可能なクラウド型音声コミュニケーションツールキットであり，学術研究用途に限り無料かつ非登録で利用可能である.

5.1 rospeex の機能

rospeex が提供する機能と想定する標準的な構成を図 5 に示す．発話理解（言語理解），対話制御，応答生成については，ユーザが記述するものとする.

rospeex では，雑音抑圧と発話区間検出はネットワーク

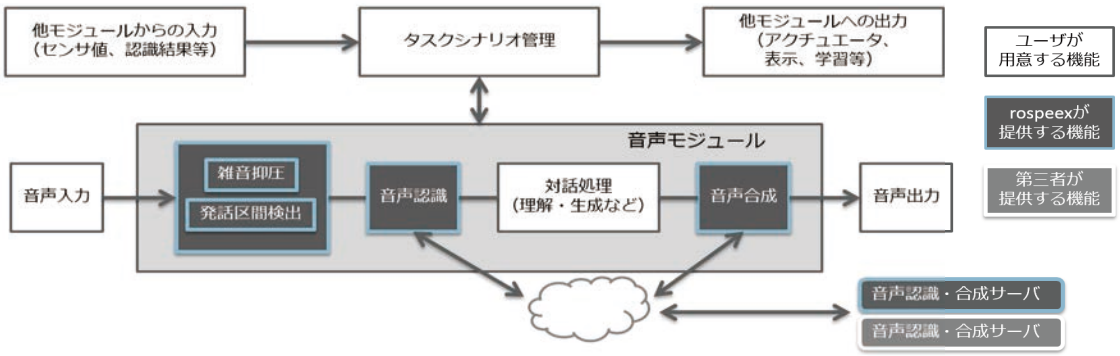


図 5: rospeex の構成の概略

上サーバで行わない設計としている．これらをサーバで処理するとネットワーク由来の遅延によりリアルタイム性の確保が難しくなるためである．また，一般的に発話区間検出の精度はそれほど高くないため，後段の処理でロボット名を含む発話のみ受け付けるなどの工夫が必要である．

rospeex は複数のクラウド型音声サービスに接続可能であり，それらを切り替えて使用できる．本節では，NICT が提供する音声認識・合成サービスについて説明する．これらは，ROS を経由せずに単体としても利用可能である．現時点では，学術研究目的に限り無償・登録不要で公開している．本サービスは，JSON ファイルをインタフェースとする．ユーザが用いるプログラミング言語には依存しないため，C++や Python など各種のプログラミング言語を利用可能である[5]．10 行程度で簡単な対話（時刻の問い合わせなど）を行う関数を記述することができる．

対話管理にマークアップ言語（VoiceXML など）を利用するソフトウェアと異なり，rospeex は対話管理の簡単なインタフェースを用意していない．これは，想定ユーザとして，複雑な対話管理を必要としないロボット開発者を念頭に置いたためである．一方，ROS 上で Python や C++ で開発したソフトウェア資産があれば，rospeex と簡単に組み合わせることが可能であるという利点がある．また，現状では，音源定位などの音響処理は統合されていない．しかしながら，HARK [Nakadai 10] など音響処理を扱うモジュールが提供されているので，rospeex の前段に容易に組み込むことが可能であると考えられる．

5.2 非モノローグ音声合成

ロボットのコミュニケーション機能の開発においては自然な音声合成が求められているが，一般的な音声合成器は人-ロボット対話に最適化されている訳ではない．rospeex では，ロボットとの対話に特化して開発されたボイスフォントを利用可能である．本ボイスフォントは，非モノローグ HMM 音声合成 [杉浦 13a] により生成される．以下では，非モノローグ HMM 音声合成について概説する．

[5]サンプルコードを `http://komeisugiura.jp/software/nm_tts.html` から入手可能である．

表 1: 学習セットの比較

システム	収録スタイル	学習セットサイズ
(0) AS	分析合成音	-
(1) Mono-176 (ベースライン)	モノローグ	176 分 (2359 文)
(2) NonM-176 (提案手法)	非モノローグ	176 分 (4485 文)
(3) NonM-325 (提案手法)	非モノローグ	325 分 (8861 文)
(4) NonM-433 (提案手法)	非モノローグ	433 分 (14179 文)

HMM の学習セットとして，声優による掛け合い対話コーパスを作成した．表 1 に示すように，声優による掛け合い対話コーパスとしては，最大級のものを用いている．サービスロボットへの応用を想定した被験者実験を行い，ベースライン手法に比べて品質が優れるという結果を得た．図 6 に提案手法とベースラインの MOS 値を示す．エラーバーは 95%信頼区間を示す．2 つの信頼区間が重なっていなければ，統計的有意差があるといえる．図より，提案手法（NonM-176，NonM-325，NonM-433）の MOS 値はベースラインと比べて高く，分析合成音（理論上の上限）に近い値であることがわかる．実験の詳細は，[杉浦 13a] を参照されたい．

非モノローグ音声合成は，ブラウザベースのサービスとして 2013 年 9 月 5 日に公開された．約半年間における音声合成サービス利用データを表 2 に示す．平均すると 1 日あたり 400 件程度の音声合成リクエストを処理している．

表 2: 実証実験の概要

実験期間	2013/9/5-2014/3/4
音声合成ユニーク IP 数	2862
音声合成リクエスト数	33320

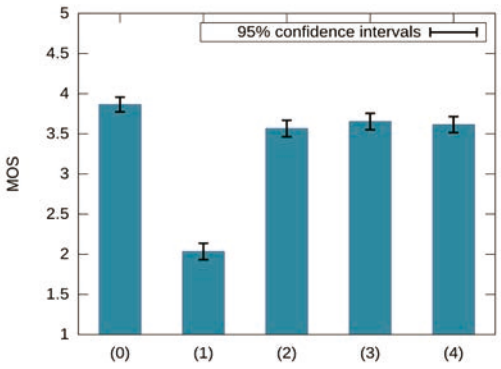


図 6: (0) 分析合成音 (理論上の上限)，(1) Mono-176 (ベースライン)，(2) NonM-176，(3) NonM-325，(4) NonM-433 に対する MOS 値.

6 おわりに

本稿では，実世界にグラウンドした音声対話に関する著者らの取り組み，日常環境における生活支援ロボットのベンチマークテストであるロボカップ@ホーム，ならびに rospeex について紹介した．高齢化社会における QOL(Quality of Life) 向上が社会的に急務であることを考えると，ロボットの日常環境への適用，ロボットによる自立支援の促進などは，ロボット開発において今後も重要な課題であろう．

将来的に家庭用多機能ロボットの普及を目標とすると，全ての機能をスタンドアロン機能として実装することはコスト面から現実的でない．一方，ネットワークへの接続を前提とすれば，音声認識や画像認識などに関する高度な技術を安価で導入することが可能である．実際に，音声での検索サービスや対話サービスの多くは，クラウド型サービスとして実装されている．現状ではクラウドロボティクスの事例は多くないものの，今後主要な分野になると考えられる．

また，ロボットに使用される技術が高度化，多機能化していくに従い，そのような複雑な機能を手軽に使用できることがますます重要になる．結果として，ユーザフレンドリなインタフェースを持つことがロボットの普及に大きな意味を持つと予想される．今日，スマートフォン上のサービスに代表されるように，適切な入出力インタフェースを選択することがユーザ体験の向上につながることは広く認識されている．ロボティクスにおいても音声と種々の入出力インタフェースをうまく統合させることが求められるようになるであろう．

参考文献

[Black 11] Black, A. W., Burger, S., Conkie, A., Hastie, H., Keizer, S., Lemon, O., Merigaud, N., Parent, G., Schubiner, G., Thomson, B., et al.: Spoken dialog challenge 2010: Comparison of live and control test results, in *Proceedings of the SIGDIAL 2011 Conference*, pp. 2–7 (2011)

[Bohus 06] Bohus, D., Langner, B., Raux, A., Black, A., Eskenazi, M., and Rudnicky, A.: Online supervised learning of non-understanding recovery policies, in *Proceedings of the IEEE/ACL Workshop on Spoken Language Technology*, pp. 170–173 (2006)

[Cohen 13] Cohen, J. and Schmidt, E.: *The New Digital Age: Reshaping the Future of People, Nations and Business*, John Murray (2013)

[Dale 95] Dale, R. and Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions, *Cognitive Science*, Vol. 19, No. 2, pp. 233–263 (1995)

[Harnad 90] Harnad, S.: The Symbol Grounding Problem, *Physica D*, Vol. 42, pp. 335–346 (1990)

[Hartanto 11] Hartanto, R.: *A Hybrid Deliberative Layer for Robotic Agents*, Springer (2011)

[Inamura 04] Inamura, T., Toshima, I., Tanie, H., and Nakamura, Y.: Embodied symbol emergence based on mimesis theory, *International Journal of Robotics Research*, Vol. 23, No. 4, pp. 363–377 (2004)

[Iocchi 10] Iocchi, L. and Zant, van der T.: RoboCup@Home: Adaptive Benchmarking of Robot Bodies and Minds, in *Proceedings of the International Conference on Simulation, Modeling and Programming for Autonomous Robots*, pp. 171–182 (2010)

[Iwahashi 09] Iwahashi, N., Taguchi, R., Sugiura, K., Funakoshi, K., and Nakano, M.: Robots that Learn to Converse: Developmental Approach to Situated Language Processing, in *Proceedings of International Symposium on Speech and Language Processing*, pp. 532–537 (2009)

[Katagiri 98] Katagiri, S., Juang, B., and Lee, C.: Pattern Recognition Using a Family of Design Algorithms based upon the Generalized Probabilistic Descent Method, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2345–2373 (1998)

[Kawahara 98] Kawahara, T., Lee, C., and Juang, B.: Flexible speech understanding based on combined key-phrase detection and verification, *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 6, pp. 558–568 (1998)

[Kehoe 13] Kehoe, B., Matsukawa, A., Candido, S., Kuffner, J., and Goldberg, K.: Cloud-Based Robot Grasping with the Google Object Recognition Engine, Proc. ICRA (2013)

[Kollar 10] Kollar, T., Tellex, S., Roy, D., and Roy, N.: Toward understanding natural language directions, in *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, pp. 259–266 (2010)

[Komatani 00] Komatani, K. and Kawahara, T.: Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output, in *Proceedings of the 18th conference on Computational Linguistics*, pp. 467–473 (2000)

[Kuffner 10] Kuffner, J.: Cloud-Enabled Robots, in *Proc. Humanoids* (2010)

[Nakadai 10] Nakadai, K., Takahashi, T., Okuno, H. G., Nakajima, H., Hasegawa, Y., and Tsujino, H.: Design and Implementation of Robot Audition System'HARK'—Open Source Software for Listening to Three Simultaneous Speakers, *Advanced Robotics*, Vol. 24, No. 5-6, pp. 739–761 (2010)

[Ogata 07] Ogata, T., Murase, M., Tani, J., Komatani, K., and Okuno, H. G.: Two-way translation of compound sentences and arm motions by recurrent neural networks, in *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and System*, pp. 1858–1863 (2007)

[Pfeifer 99] Pfeifer, R. and Scheier, C.: *Understanding Intelligence*, MIT Press, Cambridge, MA. (1999)

[Roy 02] Roy, D.: Learning visually grounded words and syntax for a scene description task, *Computer Speech and Language*, Vol. 16, No. 3, pp. 353–385 (2002)

[Stückler 12] Stückler, J., Holz, D., and Behnke, S.: RoboCup@Home: Demonstrating Everyday Manipulation Skills in RoboCup@Home, *Robotics & Automation Magazine, IEEE*, Vol. 19, No. 2, pp. 34–42 (2012)

[Sugiura 11a] Sugiura, K., Iwahashi, N., Kashioka, H., and Nakamura, S.: Learning, Generation, and Recognition of Motions by Reference-Point-Dependent Probabilistic Models, *Advanced Robotics*, Vol. 25, No. 6-7, pp. 825–848 (2011)

[Sugiura 11b] Sugiura, K., Iwahashi, N., Kawai, H., and Nakamura, S.: Situated Spoken Dialogue with Robots Using Active Learning, *Advanced Robotics*, Vol. 25, No. 17, pp. 2207–2232 (2011)

[Tenorth 12] Tenorth, M., Perzylo, A. C., Lafrenz, R., and Beetz, M.: The RoboEarth Language: Representing and Exchanging Knowledge about Actions, Objects, and Environments, in *Proc. ICRA*, pp. 1284–1289 (2012)

[Tokuda 00] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T.: Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis, in *Proceedings of ICASSP*, pp. 1315–1318 (2000)

[Yu 13] Yu, H. and Siskind, J. M.: Grounded language learning from video described with sentences, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 53–63 (2013)

[稲邑 01] 稲邑 哲也, 稲葉 雅幸, 井上 博允：ユーザとの対話に基づく段階的な行動決定モデルの獲得, 日本ロボット学会誌, Vol. 19, No. 8, pp. 983–990 (2001)

[高野 09] 高野 渉, 中村 仁彦：統計的相関に基づく動作パターンのリアルタイム教師なし分節化と原始シンボルの自律的獲得, 日本ロボット学会誌, Vol. 27, No. 9, pp. 1046–1057 (2009)

[長井 12] 長井 隆行, 中村 友昭：マルチモーダルカテゴリゼーション, 人工知能学会誌, Vol. 27, pp. 555–562 (2012)

[西田 03] 西田 豊明：人とロボットの意思疎通, 情報処理, Vol. 44, No. 12 (2003)

[松井 00] 松井 俊浩, 麻生 英樹, John, F., 浅野 太, 本村 陽一, 原 功, 栗田 多喜夫, 速水 悟, 山崎 信行：オフィス移動ロボット Jijo-2 の音声対話システム, 日本ロボット学会誌, Vol. 18, No. 2, pp. 142–149 (2000)

[山肩 04] 山肩 洋子, 河原 達也, 奥乃 博, 美濃 導彦：音声対話システムにおける物体指示のための信念ネットワークを用いた曖昧性の解消, 人工知能学会論文誌, Vol. 19, No. 1, pp. 47–56 (2004)

[岡田 10] 岡田 浩之, 大森 隆司：ロボカップ@ホーム：人とロボットの共存を目指して, 人工知能学会誌, Vol. 25, No. 2, pp. 229–236 (2010)

[河原 13] 河原達也：音声対話システムの進化と淘汰–歴史と最近の技術動向–, 人工知能学会誌, Vol. 28, No. 1, pp. 45–51 (2013)

[松田 13] 松田繁樹, 林輝昭, 葦苅豊, 志賀芳則, 柏岡秀紀, 安田圭志, 大熊英男, 内山将夫, 隅田英一郎, 河井恒, 中村哲：多言語音声翻訳システム“ VoiceTra ”の構築と実運用による大規模実証実験, 電子情報通信学会論文誌, Vol. J96-D, No. 10, p. in print (2013)

[杉浦 12] 杉浦孔明：ロボカップ@ホームリーグ, 情報処理, Vol. 53, No. 3, pp. 250–261 (2012)

[杉浦 13a] 杉浦孔明, 志賀芳則, 河井恒, 翠輝久, 堀智織：サービスロボットのための非モノローグ HMM による音声合成, 第 31 回ロボット学会学術講演会資料, pp. 2C1–02 (2013)

[杉浦 13b] 杉浦孔明, 堀智織, 是津耕司：rospeex: クラウド型音声コミュニケーションを実現する ROS 向けツールキット, 信学技報（CNR2013-10）, 第 113 巻, pp. 7–10 (2013)

[浅田 10] 浅田稔, 松原仁：ロボカップ創世記, 情報処理, Vol. 51, No. 9, pp. 1195–1200 (2010)

[大浦 13] 大浦圭一郎, 山本大介, 内匠逸, 李晃伸, 徳田恵一：キャンパスの公共空間におけるユーザ参加型双方向音声案内デジタルサイネージシステム, 人工知能学会誌, Vol. 28, No. 1, pp. 60–67 (2013)