論　文

ビッグデータの利活用による ロボットの音声コミュニケーション基盤構築

杉浦　孔明†

Analysis on Large-Scale Human-Robot Dialogues Collected by Cloud Platform

Komei SUGIURA†

あらまし　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の機能が導入され，広く一般に認知されるようになってきた．今後，ネットワークに接続されるロボットや IoT デバイスが増加するに従い，それらの機器を対象としたクラウドサービスも増加すると考えられる．一方，ロボットの音声認識・合成機能の開発は依然高コストであり，この問題を解決するためにクラウドサービスを提供することの意義は大きいと考えられる．本稿では，我々が構築した音声コミュニケーション機能を対象としたクラウドロボティクス基盤 rospeex を紹介し，長期実証実験の解析結果について報告する．

キーワード　クラウドロボティクス，音声対話システム，非モノローグ音声合成，サービスロボット

1. はじめに

スマートフォンへの音声インタフェース機能の導入は，音声言語処理機能を広く一般に認知させた．2013 年の音声認識分野の市場規模は世界全体で約 880 億円であり，2018 年には 1700 億円程度へと増加すると予測されている [1]．このような伸びの背景には，機械学習手法の進展とその大規模データへの適用とともに，音声認識機能がクラウドサービス化され，非力な端末でも高品質なサービスを利用できるようになったことがある．

一方，クラウド基盤のログとして収集される大規模なデータを用いてサービスの品質を向上することは，音声認識分野に限らず一般的に行われてきた．例えば，ユーザの購買情報から推薦精度を向上させたり，クリック数から検索精度を向上させるなどの応用は幅広く行われている [2]．これらは，検索・推薦・認識等について高機能なサービスを行うと同時に，多く利用されることでさらに性能改善を行うものである．つまり，クラウドサービスを提供することでスパイラル的改善が可能であるため，クラウドサービス提供者にもメリットがある．音声認識分野においても，クラウドサービスの提供と大規模ログコーパスを用いた精度改善が行われている [3]．

今後，ネットワークに接続されるロボットや IoT デバイスが増加するに従い，それらの機器を対象としたクラウドサービスも増加すると考えられる．実際に，クラウドロボティクス分野においては，物体認識や軌道計画などの応用について研究が始まりつつある（例えば [4]）．一方，人と共存するロボットにおいては音声対話機能の構築が高コストであり，現状では高品質なサービスが難しい．この問題を解決するためにクラウドロボティクス基盤を構築することはコミュニティへの貢献は大きいと考えられる．ただし，このような基盤構築は，音声認識・合成についての基礎技術からクラウド基盤の構築・運用やロボットへの適用に至るまでの包括的な技術開発が必要であり，簡単な課題ではない．

上記の背景から，音声対話向けクラウドロボティクス基盤 rospeex を我々は構築し，2013 年 9 月から運用してきた [5, 6](注1)．現在，4 か国語（日英中韓）の音声認識・合成に対応しており，学術研究目的に限り無

† 国立研究開発法人 情報通信研究機構 ユニバーサルコミュニケーション研究所
Universal Communication Research Institute, National Institute of Information and Communications Technology

(注1)：2016 年 1 月までに 3 万ユニークユーザに利用されている．

償・登録不要で公開している．図 1 に，rospeex のユースケースの例およびユニークユーザの分布を示す．

本稿では，rospeex の構築と長期実証実験について紹介する．本稿の構成は以下の通りである．まず，第 2 節で関連研究について述べ，次に第 3 節において rospeex を紹介する．第 4 節でクラウドロボティクス基盤を運用して得られた知見について述べたのち，第 5 節で結論を述べる．

2. 関 連 研 究

ロボットのコミュニケーション機能開発においては，無償あるいは有償のツールを用いたスタンドアロンシステムを用いるアプローチが主流であった．したがって，ロボットに搭載された計算機で処理が行われるため，CPU やメモリの制限を受ける場合がある．また，技術的には多言語対応が可能なツールが多いものの，言語モデルの置換等が必要とされるため，実際に取り組むロボット開発者は少ない．

一方，スタンドアロンのシステムではなく，音声検索，質問応答，雑談対話，音源定位・音源分離 [7] などの機能をクラウドサービスとして提供する企業や研究機関も多い．代表的なサービスとしては，Google, Microsoft，NTT docomo，Nuance などによるものが挙げられる．

また，音声以外の機能で広くネットワーク機能を利用したロボットの事例は 1990 年代から開始されている [8, 9]．また，ネットワークロボット分野では，環境中のカメラや無線タグと連携して，ロボット単体では不可能な機能を達成するアプローチとして注目を集めた [10]．近年では Kuffner によりクラウドロボティクスが提唱され，研究が盛んに行われている [11, 12]．クラウドロボティクス研究の代表事例としては，UNR-PF [10]，RoboEarth [13]，Rapyuta [14]，RoboBrain [15] などがある．

図 1　rospeex の使用例および利用者分布．左：rospeex を用いた対話の例．右：rospeex 利用者の IP アドレスの分布．

3. 音声コミュニケーションのためのクラウドロボティクス基盤 rospeex

以下では，rospeex の特徴のひとつでである非モノローグ音声合成手法 [5] について概説したのち，長期実証実験の結果得られた結果を音声認識および音声合成の面から紹介する．

3.1 非モノローグ音声合成

表現力の高い合成音声の研究開発は，音声合成分野を始めとして，音声合成の応用先である音声対話システムやロボティクスなどの分野で研究がなされてきた（例えば [16]）．我々の手法の新規性は，声優による非モノローグ収録（掛け合いで対話を行わせて収録を行う）を行った点である．また，我々が構築した非モノローグコーパスは 1 名あたり 466 分であり，[16] の約 10 倍の規模である．本コーパスの利用例としては，[17] が挙げられる．

音声合成手法として隠れマルコフモデル（HMM）に基づく方式を用いた．詳細については，[18] を参照されたい．HMM に基づく音声合成では，分析合成方式 (注2) と異なり，任意のテキストを入力として学習済みの HMM から最尤の特徴量系列を生成する．ロボティクスにおいては，動作の生成に HMM を用いる試みもある [19]．

図 2 に，読み上げタスク（左）およびサービスロボットタスク（右）において合成音声に対する品質評価を行った結果を示す．評価尺度として，標準的な Mean Opinion Score (MOS) 値を用いた．モノローグ音声合成（ベースライン）と非モノローグ音声合成手法に対して，同じ量の学習データを用いた場合の結果を比較するために，(1) と (2) を用意した．また，(0) は分析合成音（理論上の上限）であり，(3)(4) はそれぞれ (2) の 1.85 倍，2.46 倍の学習セットを用いた場合の非モノローグ音声合成手法の結果である．**および***は，$P < .01$ および $P < .001$ を示す．

図 2 より，非モノローグ音声合成とベースラインの差は読み上げタスクにおいて統計的有意でなく，非モノローグ音声合成を使用することに品質上のデメリットはないことがわかる．また，サービスロボットタスクにおいて，非モノローグ音声合成の MOS 値はベースラインと比べて高く，顕著に有効であることがわ

(注2)：分析合成方式では，音声波形を入力として，特徴量系列を直接求め，人間の発生機構を模した音源・フィルタモデルを用いて特徴量を音声波形に変換する．

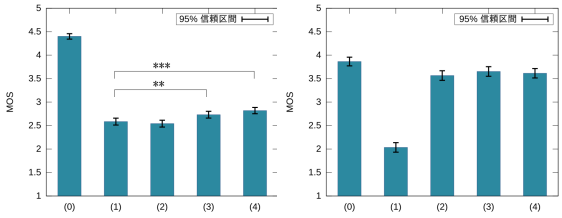


図 2　読み上げタスク（左）およびサービスロボットタスク（右）における合成音声品質評価．(1) と (2) の比較より，非モノローグ音声合成は読み上げタスクではベースラインと同等であり，サービスロボットタスクでは顕著に優れるといえる．

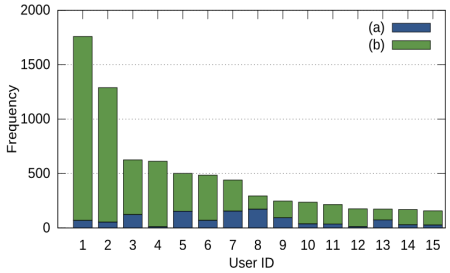


図 3　上位 15 ユーザにおける (a)N_{uniq} および (b)N_c．上位ユーザは 100〜200 種類の文を繰り返し使っているといえる．

かった．

3.2　音声認識ログの解析

2014/1/1 から 2014/11/28 までのアクセス記録をもとに，実際の利用における音声認識ログを解析した [6]．ログに含まれる発話の音声認識結果の総数は，44960 であった．ただし，無音など明らかに発話が含まれないものは取り除いた．このうち，頻度が 3 以上のものを分析対象として発話を分類した結果を表 1 に示す．音声認識結果を以下のカテゴリに分類した．

(a)　挨拶・雑談: 日常会話（例：こんにちは）

(b)　一問一答型質問: 対話履歴を必要としない情報源への問い合わせ（例：今何時）

(c)　移動・把持: 移動や把持に関連する動作指示発話（例：止まれ）

(d)　家電操作: 音声リモコンのように家電を操作する発話（例：テレビを消して）

(e)　認識・学習: センサ入力の学習または認識を指示する発話（例：ここはどこ）

(f)　一般的な指示: (c)-(e) 以外でロボットの行動を指示する発話（例：手を上げろ）

(g)　その他（検索・回答，判別不能）: (a)-(f) 以外の発話．主に，質問への応答，音声認識誤りまたは判別不能な発話を含む．

表 1 より，約半数の発話は挨拶・雑談や一問一答型の質問であったことがわかる．これらの発話に対しては，一般的に提供されている質問応答や雑談対話のクラウドサービスを用いることが有効であると考えられる．

表 1　発話の分類

カテゴリ	発話数	割合 [%]
挨拶・雑談	1894	31.59
一問一答型質問	1153	19.23
移動・把持	258	4.30
家電操作	229	3.82
認識・学習	215	3.59
一般的な指示	41	0.68
その他（検索・回答，判別不能）	2205	36.78
合計	5995	100

3.3　音声合成に関する個人依存性

これまでロボティクスにおいてはスタンドアロンアプローチが採られてきたため，ロボット開発者が求める音声処理機能の性質について，多様かつ多人数のユーザを扱った定量的な解析が困難であった．クラウドロボティクス基盤を構築することで，ロボット対話開発における個人依存性を調査することが初めて可能になる [20]．

いま，あるユーザ u_i の音声合成リクエスト履歴全体を文集合 S_i と定義する．S_i に含まれる文の種類を N_{uniq} とすると，S_i 中でキャッシュを利用できるリクエスト数 N_c は，$N_c = ||S_i|| - N_{uniq}$ と表すことができる．ここで，擬似キャッシュヒット率 r_p を $r_p = \frac{N_c}{||S_i||}$ と定義する．r_p が高いユーザは，同じ言い回しを繰り返し使っていると考えられる．

2014/1/1 から 2015/5/31 までの rospeex のユーザのうち，10 回以上合成リクエストを行ったユーザを抽出した．そのうち，上位 15 ユーザについて解析した結果を図 3 に示す．図において，(a) は文の種類 N_{uniq}，(b) は N_c を表す．N_{uniq} は N_c に比べて小さく，上位ユーザは 100〜200 種類の文を繰り返し使っていることがわかる．実験の詳細については，[20] を参照されたい．

4.　クラウドロボティクス基盤運用に関する留意点

4.1　クラウドサービス提供者のメリット・デメリット

社会展開の手段としてソフトウェアの公開は広く行われているが，データセットや学習済みのモデルが知

財として高い価値を持つ場合，それらが無償で公開されることは稀である．一方，クラウドサービスであれば，データセットを公開する必要なく社会展開が可能である．また，ソフトウェアの影響を計測したい場合，ダウンロード数では実際に使用したユーザ数を反映していない可能性がある．一方，クラウドサービスではサーバ上へのアクセス数からユーザが使用した回数を計測することができる．

一方，クラウドサービス提供にはデメリットも存在する．クラウドサービスの知名度が高くなると，サービスそのものやウェブサイトが攻撃や不正利用の対象になりやすい．また，ロボットの機能が全てクラウド化されるべきなのではなく，リアルタイムの衝突回避など，クラウド化が容認されにくい機能も存在するので，全てのクラウドサービスがユーザを獲得できるわけではない．ユーザ数の増加が見込めないようであれば，クラウドサービス化以外の方法が費用対効果の面で有利になる場合もある．

4.2 データの共有

1. 節で述べたように，情報検索や推薦分野では研究促進を目的として，購買履歴等のデータセットが公開されている．同様に，クラウドロボティクス基盤で収集されたデータをコミュニティで共有するためには，どのような議論点が存在するのであろうか？

クラウド型音声検索サービスを利用したユーザが「暗証番号は xxxx です」という発話を入力する可能性は零ではない．さらに，認識結果には誤りが含まれるので，認識結果の文字列マッチングで不適切な発話を除外することも難しい．すなわち，全ての発話について問題がないか人手で確認する作業が必要になる．

ユーザからの視点としては，サービス利用で得られる便益とのバランスによりプライバシー情報を提供するかどうかを決定できることが重要である．これは，スマートフォンにおいて位置情報等のプライバシー情報の利用許諾が事前に設定できるようになっている仕組みと同様である．そのためには，クラウドへアップロード済みのデータについても，ユーザが消去できる機能を用意するなどの対策が考えられる．実際に，Google のサービスではユーザ ID と紐付けられた音声をユーザ自身が確認し消去可能である．ただし，サービス開始時からこの機能が提供されていた訳ではなく，全てのケースでこのような機能を導入することは実際には難しい点に注意が必要である．音声に限定しないネットワークロボット技術の法的側面については [21] が詳しい．

5. おわりに

クラウドロボティクス技術は屋内レベルから屋外，都市に至るまで多岐に亘る応用を持つ．代表的な応用対象としては，スマートホームデバイス，ドローン，プローブカーなどが挙げられる．本稿では，主にサービスロボットの音声コミュニケーション機能を対象としたクラウドロボティクス基盤 rospeex について述べた．

本研究に関する動画や，rospeex のインストール方法は `http://rospeex.org` を参照されたい．

文　献

[1] NEDO 技術戦略研究センター，“Tsc foresight vol. 8,” 2015.

[2] D. Jannach，M. Zanker，A. Felfernig, and G. Friedrich, 情報推薦システム入門 -理論と実践-，共立出版，2012.

[3] 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，pp.2549–2561，2013.

[4] D. Berenson, P. Abbeel, and K. Goldberg, “A robot path planning framework that learns from experience,” IEEE ICRA, pp.3671–3678, 2012.

[5] K. Sugiura, Y. Shiga, H. Kawai, T. Misu, and C. Hori, “A Cloud Robotics Approach towards Dialogue-Oriented Robot Speech,” Advanced Robotics, vol.29, no.7, pp.449–456, 2015.

[6] K. Sugiura and K. Zettsu, “Rospeex: A cloud robotics platform for human-robot spoken dialogues,” Proc. IEEE/RSJ IROS, pp.6155–6160, 2015.

[7] 水本武志，中臺一博，“Hark saas:ロボット聴覚ソフトウェア hark のクラウドサービスの設計と開発,” 人工知能学会研究会資料，pp.60–65，2015.

[8] M. Inaba, “Remote-Brained Robotics: Interfacing Al with Real World Behaviors,” Proc. of the 6th Int. Symp. of Robotics, pp.335–344, 1993.

[9] K. Goldberg, M. Mascha, S. Gentner, N. Rothenberg, C. Sutter, and J. Wiegley, “Desktop teleoperation via the world wide web,” Proc. IEEE ICRA, pp.654–659, 1995.

[10] K. Kamei, S. Nishio, N. Hagita, and M. Sato, “Cloud Networked Robotics,” Network, IEEE, vol.26, no.3, pp.28–34, 2012.

[11] J. Kuffner, “Cloud-Enabled Robots,” Proc. Humanoids, pp.●●–●●, 2010.

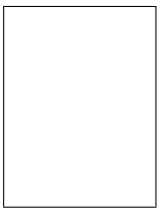
[12] B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg, “A survey of research on cloud robotics and automation,” IEEE Trans. on Automation Science and Engineering, vol.12, no.2, pp.398–409, 2015.

[13] M. Tenorth, A.C. Perzylo, R. Lafrenz, and M. Beetz, “The RoboEarth Language: Representing and Exchanging Knowledge about Actions, Objects, and Environments,” Proc. ICRA, pp.1284–1289, 2012.

[14] G. Mohanarajah, D. Hunziker, R. D’Andrea, and M. Waibel, “Rapyuta: A cloud robotics platform,” IEEE Trans. Automation

Science and Engineering, vol.12, no.2, pp.481–493, 2015.
[15] A. Saxena, A. Jain, O. Sener, A. Jami, D.K. Misra, and H.S. Koppula, “Robobrain: Large-scale knowledge engine for robots,” 2014.
[16] K. Iwata and T. Kobayashi, “Conversational Speech Synthesis System with Communication Situation Dependent HMMs,” Proc. of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop, pp.113–123, 2011.
[17] T. Nose, Y. Arao, T. Kobayashi, K. Sugiura, Y. Shiga, and A. Ito, “Entropy-based sentence selection for speech synthesis using phonetic and prosodic contexts,” Proc. INTERSPEECH, pp.3491–3495, 2015.
[18] 志賀芳則，河井恒，“多言語音声合成システム,” 情報通信研究機構季報，vol.58，no.3，pp.19–24，2012.
[19] K. Sugiura, N. Iwahashi, and H. Kashioka, “Motion Generation by Reference-Point-Dependent Trajectory HMMs,” Proc. IROS, pp.350–356, 2011.
[20] 杉浦孔明，是津耕司，“クラウドロボティクス基盤 rospeex の長期実証実験と大規模ロボット対話データの解析,” 第 33 回日本ロボット学会学術講演会資料集，pp.RSJ2015AC3D1–03，2015.
[21] 土井美和子，小林正啓，萩田紀博，ユビキタス技術 ネットワークロボット - 技術と法的問題，オーム社，2007.

（平成 xx 年 xx 月 xx 日受付）



杉浦　孔明

2002 年京大・工・電気電子工学科卒．2007 年同大学院・博士課程了．日本学術振興会特別研究員，ATR 音声言語コミュニケーション研究所研究員を経て，現在，情報通信研究機構ユニバーサルコミュニケーション研究所主任研究員．サービスロボット，機械学習，クラウドロボティクス，音声対話システム，情報検索，推薦システムなどの研究に従事．計測自動制御学会，情報処理学会，人工知能学会，日本ロボット学会，IEEE などの会員．