様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 １ 日現在

機関番号：１４５０１
研究種目：研究活動スタート支援
研究期間：2019～2020
課題番号：１９Ｋ２４３４３

研究課題名（和文）サービスロボットのための雑音に頑健な音声認識および音声対話の研究

研究課題名（英文）Noise-robust speech recognition and spoken dialog system for service robots

研究代表者
高島　遼一（Takashima, Ryoichi）

神戸大学・都市安全研究センター・准教授

研究者番号：５０８４６１０２

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究では、音声対話の分野において従来独立に最適化されていた音声認識や対話のモジュールを一体化し、音声対話成功という損失関数の元で学習するEnd-to-Endモデルの構築を最終目的としている。しかし一般にこのモデルの学習には膨大な学習データが必要であるため、限られたデータ量でモデルを学習するための手法の開発が必要である。そこで本研究では、音声認識と対話のモデルに対して限られたデータ量であっても安定した学習を行うための手法として、多段階の転移学習や自己教師あり学習、外部知識の利用といった手法を提案し、音声認識、対話タスクにおいて従来法よりも性能の高いモデルを学習可能であることを確認した。

研究成果の学術的意義や社会的意義
近年の労働力不足の問題解決のため、サービスロボットに対するニーズが高まっている。音声によるロボットとの対話はユーザにとって馴染みやすいが、高雑音環境といった音声認識が困難な状況では期待した対話性能が得られない。従来、このような問題に対して音声認識、対話技術が個別に最適化される形で研究されており、必ずしも音声対話成功という最終目的に対して最適化がされていなかった。これらのモジュールを一本化して全体最適化が行えればさらに性能向上が見込まれるが、これには膨大な学習データが必要である。本研究の成果は、限られた学習データで安定してモデルを学習する方式であり、前述の全体最適化に利用可能と期待している。

研究成果の概要（英文）：The final of this research is to construct an End-to-End model which handles both speech recognition and dialogue system. In the field of speech dialogue system, the conventional system independently optimizes speech recognition module and dialogue module. However, the training of End-to-End model requires huge training data; therefore, the technique to train models on limited training data is important. For this reason, in this research, we propose training techniques using multi-step transfer learning, self-supervised learning, and external knowledge, and confirm that our proposed method can training models showing better performance than conventional methods.

研究分野：メディア情報処理

キーワード：音声認識　音声対話　ニューラルネットワーク　機械学習

１．研究開始当初の背景

少子高齢化や訪日観光客の増加といった社会現象に伴い、サービス業における労働力不足が加速している。2018 年時点でのサービス業の人手不足数が約 46 万人であった[1]のに対して、2030 年には 400 万人にも達すると予想されている[2]。このような背景から、サービスロボットに対するニーズが高まっており、その市場規模は 2020 年で約 1 兆円、2035 年には 4.9 兆円まで成長することが見込まれている[3]。

サービスロボットとユーザとのインタラクション方法として、音声による対話は人間にとってなじみが深く、最も直感的な方法の一つであると言える。AI スピーカのような音声入力を用いたデバイスが一般家庭用に普及している。しかし家庭内で比較的定型のフレーズを入力する AI スピーカと異なり、サービスロボットが使われる環境は商業施設や公共施設といった雑音の多い環境が多く、また複数の相手と同時に対話するケースやロボットから距離の離れたユーザと対話するケースも存在し、さらに話し口調の音声を認識する必要があるため、音声認識としては最も難しいタスクの一つと言える。そしてこのような誤りやすい音声認識結果から、ロボットはユーザの期待する返答を推定しなければならない。そのため、サービスロボットとの音声による対話はユーザにとって馴染みやすいものの、期待した性能が得られていないというのが現状である。

[1] 厚生労働省,“平成 30 年上半期雇用動向調査結果”
[2] パーソル総合研究所・中央大学,“労働市場の未来推計 2030”
[3] 経済産業省・NEDO,“平成 22 年ロボット産業将来市場調査”

２．研究の目的

本研究では、サービスロボットとの実用的な音声対話のための技術について研究を行う。具体的には、高雑音・複数話者・遠隔音声・話し言葉音声に頑健な音声認識技術、および音声認識性能が低い状況に頑健な音声対話（応答文生成と音声合成）の技術確立を目的とする。

外乱に頑健な音声認識の研究、また音声対話の従来研究は多く存在する。従来の研究では図１左図のように、雑音除去、音声認識、応答文生成、音声合成といった機能がそれぞれ独立のモジュールとして、個別の目的関数により最適化されている。例えば雑音除去では信号雑音比（SN 比）、音声認識では音声認識率を目的関数とし、それらを最大化するように個別最適化している。しかし個別最適化は、各目的関数が最終的な目的(本研究ではロボットが適切な応答を返すこと)に対して間接的であり、最終目的に対して最適化がされるとは限らない。例えば人間は、話し相手の声の一部が雑音で聞き取りにくかったとしても、前後の文脈から何と話したのかが推定でき、また一部の助詞が聞き取れなかったとしても、重要な単語さえ聞き取れれば会話を成立させることが可能である。これは前述の例で言い換えると、前者は、音声認識にとって全ての時間帯で SN 比を最大にすることは必須では無く、また後者は、音声対話にとって全ての単語の認識を正解することは必須では無いという可能性を示唆している。そこで本研究では、近年盛んに研究が進んでいる深層学習をベースに、音声入力から応答までの全モジュールを異なる目的関数による個別最適化ではなく、図１右図のように対話成功率という最終的な目的関数の上で全体最適化することで、従来の音声対話システムの性能を向上させることを最終目標とする。

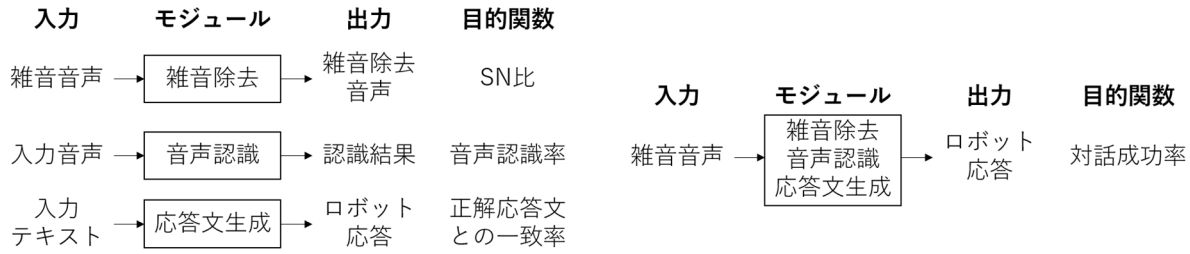

図 1.　従来のモジュール個別最適化(左図)と目的とする全体最適化(右図)の各例

３．研究の方法

前述のとおり、最終的な目標は音声入力から応答文生成までを全体最適化することにより、実環境音声の音声対話の成功率を実用化レベルまで向上させることである。しかし一般に、複数のモジュールを単一のネットワークで表現する End-to-End モデルの学習には膨大なデータが必要であるため、利用可能な学習データでは安定したモデル学習が行えないことが予想される。そこで本研究では最終目標に対するサブ目標として、音声認識までのモジュールと、対話のモジュールで切り分け、２モジュールそれぞれを効率よく学習させることを検討した。

２モジュールに切り分けたとしても、それぞれを安定して学習させるには膨大なデータが必要である。例えば音声認識モジュールについては、従来の音声認識モジュールはニューラルネットワークモデルや隠れマルコフモデル、発音辞書、言語モデルといった複数のモジュールで構成

されており、これらは独立に学習されている。本研究ではこれらを単一のニューラルネットワークモデルで表現しているため、従来モデルと比較して大量の学習データが必要となる。そこで、本研究では、Transfer learning や Self-supervised learning といったアプローチを用いることで、目的環境の学習データが限られている場合であっても安定してモデルを学習することを検討した。また対話に関しても同様に、限られたデータ量でも安定したモデル学習を行うために、外部知識を利用した学習方法を検討した。

４．研究成果

本節では、音声認識モデル、対話モデルに関する主な研究成果を述べる。

【音声認識モデルに関する研究成果】

Transfer learning（転移学習）を応用したモデル学習手法として、多段階転移学習と呼ぶ手法を開発した[4]。転移学習とは、目的環境の学習データが少ない状況において、まず目的環境以外の大量の学習データを用いてモデルを事前学習しておき、その後目的環境の学習データを用いて fine-tuning する手法である。

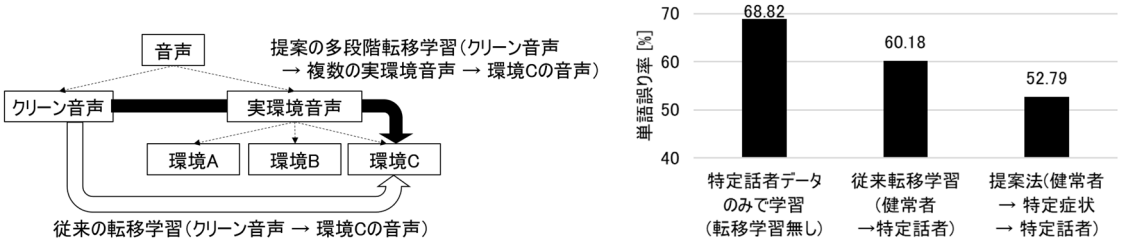

図２．提案の多段階転移学習の概要(左)と、構音障害者音声認識タスクでの評価結果(右)

提案手法の概要を図 2 の左に示す。従来の転移学習では、まず一般的な音声データベース（雑音の無いクリーンな音声が多い）を用いてモデルを学習し、その後目的環境（図の例では環境 C）の音声データを用いて fine-tuning する。一方提案法では、環境 A や B のデータも混ぜた「一般的な実環境音声」を用いて fine-tuning した後に環境 C のデータでさらに fine-tuning を行う。これにより、クリーン音声環境と目的環境のデータの性質が大きく異なる場合において、段階を踏んでモデルを目的環境へ適応することが可能となり、安定したモデル学習が可能となる。雑音環境ではないが、似たタスクとして構音障害者の音声認識タスクを用いて本手法を評価したところ、健常者で事前学習して特定障害者へ fine-tuning する従来法に比べて、一旦複数の障害者へ fine-tuning してから特定障害者へ fine-tuning する提案法の方が、音声認識誤り率が小さくなることが明らかとなった（図 2 右）。

Self-supervised learning を用いたモデル学習手法として、Self-supervised learning と Transfer learning を組み合わせた手法も開発した[5]。Self-supervised learning とは、ラベル（書き起こしテキスト）の無い大量の音声データを用いてモデルの事前学習を行う手法である。

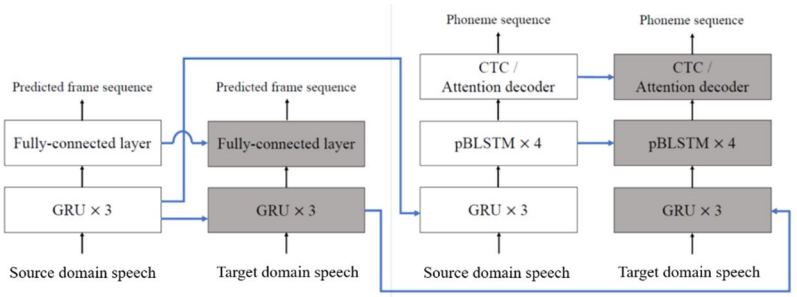

図３．Self-supervised learning と Transfer learning の組み合わせによるモデル学習法

Self-supervised learning と Transfer learning を組み合わせる場合、様々な学習手順が考えられるが、我々が検討した結果、図 3 のような手順で学習するのが最も音声認識性能が良いということが分かった。この手法も構音障害者音声認識タスクで評価した結果、Self-supervised learning も Transfer learning も用いないベースラインの音声認識誤り率が 29.9%であったのに対して、開発手法を用いた場合では 13.7%の大幅な性能改善を示した。

【対話モデルに関する研究成果】

対話モデルの学習を難しくする要因の一つとして、単語の多様性に比べて学習データが十分でないことが挙げられる。例えばサッカーのような単語は学習データに多く存在するため、サッカーが入力された場合においては比較的正しい応答が出力可能であるが、学習データに少ししか存在しないスポーツの場合、正しい応答を出力するのが困難である。このとき、この学習

データの少ない単語が、サッカーと同じスポーツであるという外部知識があれば、正しい応答ができると期待される。そこで、WordNet と呼ばれる単語の類義関係を記したデータベースである WordNet を外部知識として用いる手法を提案した[6]。

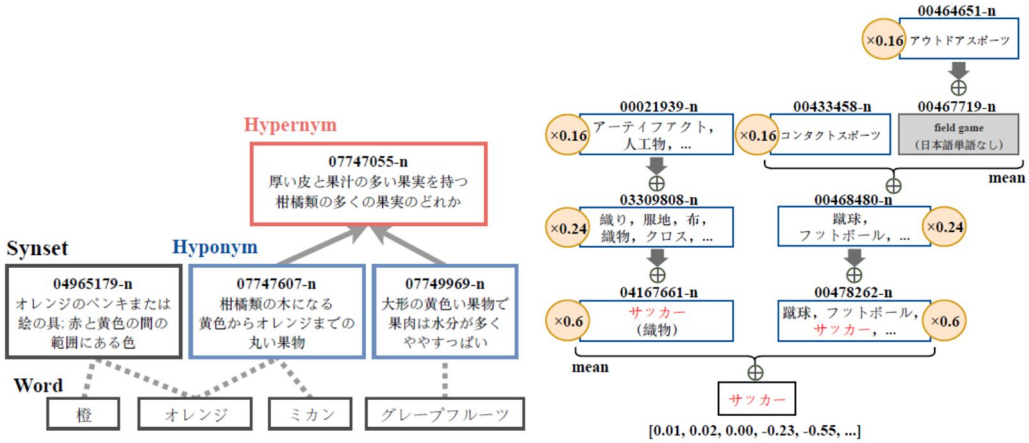

図 4．WordNet（左）と提案する入力ベクトル作成手法の概要（右）

WordNet は図 4 の左に示されるような、各単語に対して語義や概念でグループ化されたネットワーク構造である。例えば橙、オレンジ、ミカン、グレープフルーツといった単語が従来では全く別の単語として定義されていたのに対して、WordNet を用いることで、これらの関係も併せてモデルを学習することが可能となるため、学習データが少ない単語の問題が軽減されると考えられる。提案手法を図 4 の右に示す。提案手法では、例えばサッカーという単語をベクトル表現にしてモデルへ入力する際に、アウトドアスポーツなどといった情報のベクトル表現にして加算した上で入力する。

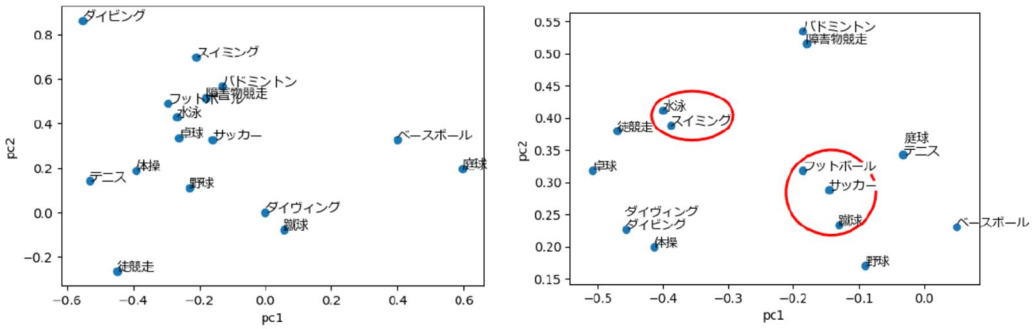

図 5.従来手法（左）と提案手法（右）における単語表現の散布図

従来手法でベクトル表現した単語と、提案手法でベクトル表現した単語それぞれの 2 次元プロットを図 5 に示す。提案手法を用いた場合、「水泳とスイミング」、「フットボールとサッカーと蹴球」が似た表現になっていることから、単語の関係がうまくモデル化されていることが分かる。これにより、応答出力の評価値も向上（BLUE-1、BLUE-2 でそれぞれ 1.74%と 7.55%の相対改善）できることを確認している。

以上のことから、音声認識モジュール、対話モジュールそれぞれについて、限られた学習データにおいて性能の高いモデルを学習可能な手法を提案した。今後はそれぞれ学習された二つのモジュールを統合して、さらに最終目的である対話成功率基準での End-to-End モデルの構築を目指していく。

(発表文献)

[4] Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki,
Two-Step Acoustic Model Adaptation for Dysarthric Speech Recognition,
ICASSP, pp. 6104-6108, May 2020.

[5] 澤 佑哉, 冨士原 健斗, 相原 龍, 高島 遼一, 滝口 哲也, 本山 信明
自己教師あり学習によるラベル無し自由発話を用いた構音障害者音声認識,
日本音響学会 2021 年春季研究発表会講演論文集, 2-2P-2, pp. 1045-1048, 2021-03.

[6] 麻生大聖, 高島遼一, 滝口哲也, 有木康雄,
WordNet を用いた雑談対話システムの汎化性能の向上,
電子情報通信学会技術研究報告, Vol. 119, No. 188, pp. 19-24, 2019-08.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Yuki Takashima, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki	４．巻 7
２．論文標題 Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition	５．発行年 2019年
３．雑誌名 IEEE Access	６．最初と最後の頁 164320-164326
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2019.2951856	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計20件（うち招待講演　0件／うち国際学会　5件）

１．発表者名 Yuki Takashima, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki
２．発表標題 Dysarthric Speech Recognition Based on Deep Metric Learning
３．学会等名 Interspeech（国際学会）
４．発表年 2020年

１．発表者名 Weihao Zhuang, Tristan Hascoet, Ryoichi Takashima, Tetsuya Takiguchi and Yasuo Ariki
２．発表標題 Convolutional neural networks Memory Optimization Inference with Splitting Image
３．学会等名 IEEE Global Conference on Consumer Electronics (GCCE)（国際学会）
４．発表年 2020年

１．発表者名 Yuya Sawa, Ryoichi Takashima, Tetsuya Takiguchi
２．発表標題 An Investigation of End-to-End Speech Recognition Using Model Adaptation for Dysarthric Speakers
３．学会等名 IEEE Global Conference on Consumer Electronics (GCCE)（国際学会）
４．発表年 2020年

１．発表者名
冨士原 健斗，高島 遼一，杉山 千尋，田中 信和，野原 幹司，野崎 一徳，滝口 哲也

２．発表標題
口唇口蓋裂者の音声認識のためのデータ拡張方式の検討

３．学会等名
日本音響学会2021年春季研究発表会講演論文集

４．発表年
2021年

１．発表者名
陳 訓泉，陳 金輝，高島 遼一，滝口 哲也

２．発表標題
Dysarthric Speech Conversion by Learning Disentangled Representations with Non-parallel Data

３．学会等名
日本音響学会2021年春季研究発表会講演論文集

４．発表年
2021年

１．発表者名
澤 佑哉，冨士原 健斗，相原 龍，高島 遼一，滝口 哲也，本山 信明

２．発表標題
自己教師あり学習によるラベル無し自由発話を用いた構音障害者音声認識

３．学会等名
日本音響学会2021年春季研究発表会講演論文集

４．発表年
2021年

１．発表者名
澤 佑哉，高島 遼一，滝口 哲也，有木 康雄

２．発表標題
構音障害者音声認識における発話辞書適応の検討

３．学会等名
日本音響学会2020年秋季研究発表会講演論文集

４．発表年
2020年

１．発表者名
Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki

２．発表標題
Two-step acoustic model adaptation for dysarthric speech recognition

３．学会等名
2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)（国際学会）

４．発表年
2020年

１．発表者名
Weihao Zhuang, Hascoet Tristan, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki

２．発表標題
Optimizing the Computational Efficiency of 3D Segmentation Models for Connectomics

３．学会等名
The 26th International Workshop on Frontiers of Computer Vision (IW-FCV 2020)（国際学会）

４．発表年
2020年

１．発表者名
澤佑哉，高島遼一，滝口哲也，有木康雄

２．発表標題
Hybrid CTC/attentionモデルを用いた構音障害者音声認識の検討

３．学会等名
日本音響学会2020年春季研究発表会講演論文集

４．発表年
2020年

１．発表者名
南坂竜翔，高島遼一，滝口哲也

２．発表標題
少量データを用いた構音障害者音声合成の健常者モデルによる明瞭性改善

３．学会等名
日本音響学会2020年春季研究発表会講演論文集

４．発表年
2020年

１．発表者名
黄伊莎，Tristan Hascoet，高島遼一，滝口哲也，有木康雄

２．発表標題
Differentiable Programmingを用いた強化学習の最適化

３．学会等名
情報処理学会第82回全国大会講演論文集

４．発表年
2020年

１．発表者名
長谷川貴大，Tristan Hascoet，高島遼一，滝口哲也，有木康雄

２．発表標題
ニューロンセグメンテーションにおけるマルチドメイン学習による汎化性能の改善

３．学会等名
情報処理学会第82回全国大会講演論文集

４．発表年
2020年

１．発表者名
高島悠樹，高島遼一，滝口哲也，有木康雄

２．発表標題
構音障害者音声認識のための健常者音声及び他言語障害者音声を用いた転移学習

３．学会等名
電子情報通信学会技術研究報告

４．発表年
2019年

１．発表者名
麻生大聖，高島遼一，滝口哲也，有木康雄

２．発表標題
外部知識を用いた雑談対話システムの汎化性能向上の検討

３．学会等名
日本音響学会2019年秋季研究発表会講演論文集

４．発表年
2019年

１．発表者名
Zhaojie Luo, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki

２．発表標題
Speech-to-Speech Translation using Dual Learning and Prosody Conversion

３．学会等名
日本音響学会2019年秋季研究発表会講演論文集

４．発表年
2019年

１．発表者名
南坂竜翔，高島遼一，滝口哲也，有木康雄

２．発表標題
構音障害者の少量データを用いた深層学習による音声合成の検討

３．学会等名
日本音響学会2019年秋季研究発表会講演論文集

４．発表年
2019年

１．発表者名
高島遼一，滝口哲也，有木康雄

２．発表標題
構音障害者を対象とした日本語大語彙連続音声認識の検討

３．学会等名
日本音響学会2019年秋季研究発表会講演論文集

４．発表年
2019年

１．発表者名
麻生大聖，高島遼一，滝口哲也，有木康雄

２．発表標題
WordNetを用いた雑談対話システムの汎化性能の向上

３．学会等名
電子情報通信学会技術研究報告

４．発表年
2019年

１．発表者名
Weihao Zhuang, Tristan Hascoet, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki

２．発表標題
Reduce GPU Memory Usage of Training Neural Network by CPU Offloading

３．学会等名
第22回画像の認識・理解シンポジウム

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
研究者webページ
http://www.me.cs.scitec.kobe-u.ac.jp/~rtakashima/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関