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

平成　３０　年　　６　月　１５　日現在

機関番号：　１２６０８
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2017
課題番号：　２６２８００５５

研究課題名（和文）音声認識生成システムの自己組織化学習

研究課題名（英文）Self-Organized Learning of Speech Recognition and Synthesis Systems

研究代表者
篠崎　隆宏（Shinozaki, Takahiro）

東京工業大学・工学院・准教授

研究者番号：８０４４７９０３

交付決定額（研究期間全体）：（直接経費）　　12,100,000円

研究成果の概要（和文）：少量のラベル付き音声データと大量のラベル無し音声データから音声言語を学習したり、人手に頼ることなく自動的にシステム構造や学習条件を最適化しシステム性能を最大化したりすることのできる、自律的な音声言語情報処理システムの仕組みを実現することを目的に研究を行った。進化戦略を用いた大規模なニューラルネットワークシステムの自動最適化手法や、音声をモデル化する各種の統計モデルの教師なし学習法、強化学習法の提案を行い、実験により有効性を示した。研究成果発表の一環として公開したフリーな高性能日本語音声認識システムは、国内外で幅広く用いられている。

研究成果の概要（英文）：The purpose of this study is to make self-standing speech and language information processing systems that can learn from a small amount of labeled and a significant amount of unlabeled speech data as well as can automatically optimize its structure and learning conditions. We have proposed evolution strategy based automation method for neural network-based system development, series of semi-supervised learning methods for statistical speech models, and a reinforcement learning method of speech recognition systems. A high-performance Japanese speech recognition system integrating the research results have been published and widely used.

研究分野：　音声言語情報処理

キーワード：　音声言語情報処理　深層学習　モデル構造最適化　半教師あり学習　高性能音声認識システム　強化学習　ブラックボックス最適化　進化的アルゴリズム

１．研究開始当初の背景

音声認識や音声合成などの高度な音声情報処理システムは、大規模な統計モデルを書き起こしラベル付きの音声データを大量に用いて教師あり学習することにより実現されている。しかしビッグデータの時代と言われる今日であっても、音声に対応した書き起こしテキストは人手で作成せざるを得ない。また、そのようなデータを集めたとしても、実際に高い性能を達成するにはシステムを人手により高度にチューニングする必要がある。このため高性能な音声言語処理システムの開発は非常に大きな手間と費用のかかるものとなっており、幅広い応用を行う上での足かせとなっている。

２．研究の目的

少量のラベル付き音声データと大量のラベル無しデータから音声言語を学習したり、人手に頼ることなく自動的にシステム構造や学習条件を最適化し性能を最大化したりすることのできる、自律的な音声言語情報処理システムの仕組みを実現することを目的とする。

３．研究の方法

本研究では、ニューラルネットを用いて実装した音声認識システムや音声合成システム等の音声言語情報処理システムを対象とする。少量のラベル付き音声データと大量のラベル無し音声データをもとに音声モデルの半教師あり学習を実現する方法について研究し、様々なタスクにおいて評価実験を行う。さらに、ニューラルネットワークの構造や学習条件を自動的に調整しシステム性能を最大化するため、ブラックボックス最適化手法を応用した仕組みについて研究する。音声情報処理で用いられるニューラルネットワークは大規模なものであり、自動最適化においてはクラウド計算機を前提に効率的な並列計算が可能なソフトウエアを開発し、実験に用いる。

４．研究成果

ニューラルネットワークを用いた音声言語情報処理システムのネットワーク構造や学習係数等の学習メタパラメタの自動最適化は、メタパラメタ集合を入力、システム性能を出力とするブラックボックス関数の最適化問題と捉えることができる。ポイントとなるのは、関数形が複雑またはその他の理由で解析的な最適化は不可能であるものの、入力が与えられれば出力を求めることは可能であることである。例えば音声認識システムを構成するニューラルネットワークの構造設計に対して評価用音声における認識誤り率を評価しようとする場合、ネットワーク構造と認識誤り率の関係は複雑であり簡潔な数式で表すことはできない。しかし、実際に音声認識システムを構築して認識実験を行えば、認識誤り率を評価することはできる。もしメタパラメタ数が１個か２個程度であれば、グリッドサーチで一通りの値の組み合わせを試すことで、メタパラメタの最適値を探索できる。しかし、メタパラメタの数が少し増えただけでそれらの値の組み合わせは指数関数的に増加するため、この方法はすぐに破綻してしまう。そこで、ベイズ的最適化や遺伝的アルゴリズム、進化戦略の適用を検討した。

ベイズ的最適化は、関数を確率変数として扱い、ベイズ推論の枠組みを用いて解候補の探索を逐次的に行う手法である。すなわち第一にブラックボックス関数の事前分布を仮定し、既知の観測サンプル集合をもとに予測分布を求める。観測サンプルは、ここでは学習メタパラメタとシステム性能の対である。そして第二に、得られた予測分布をもとに次の一手の良し悪しを評価する獲得関数を最大化するよう解候補となる次の探索点を決める。この操作を繰り返すことで、最適解を得ることを目指す。遺伝的アルゴリズムは、自然界での生物の進化を模倣したアルゴリズムである。解の候補を整数等の数値の並びで表現したものを遺伝子とみなし、遺伝子に対応した個体の評価をもとに遺伝子集合を逐次的に更新する。遺伝子の更新は世代交代になぞらえられ、優良個体の選択や複数の親からの遺伝子の組み換え、および突然変異の組み合わせにより、新しい遺伝子を生成していく。世代を重ねるごとにより優れた個体が多く出現することを期待する。進化戦略は遺伝的アルゴリズムと類似した手法であるが、遺伝子を固定長の実数ベクトルで表現し問題を定式化するのが特徴である。中でも共分散行列適応進化戦略(CMA-ES)は、様々なブラックボックス最適化に有効であることが示されている。

音声言語情報処理システムのブラックボックス最適化にあたり、ニューロンレイヤーが一列に並んだ一般的な構成のニューラルネットワークの他、ノードをニューロンレイヤーとする任意の有向無サイクルグラフ構造を考え、最適解探索の対象とした。グラフ構造を遺伝子として数値列にエンコードする方法として、任意の有向グラフのノードはすべての有向枝が一定の方向を向くように一列に並べることができることを用い、三角隣接行列としてレイヤー間の結線構造を表現し、さらにその要素を一列に並べ替える手法を提案した。遺伝子には他に各レイヤーのニューロン数や学習率等の学習メタパラメタも含めることが可能であり、ネットワーク構造と各種学習条件が一つの数値ベクトルとして表現される。ブラックボックス最適化手法として遺伝的アルゴリズムを用いる場合はベクトルの各次元の値として整数や実数等を使い分けることも可能であるが、進化戦略を用いる場合は遺伝子のすべての要素は実数である必要がある。そこで、各メタパ

ラメタの特性に応じて任意の実数値を正の実数や自然数などに対応させる関数を事前に設定した。

スーパーコンピュータ上で単語検出器や音声認識システム、機械翻訳システム等を自動最適化する実験を行った。CMA-ES が他の手法と比較して効率的な最適化が可能であるとともに利用も容易であることを示した。ニューラルネットの学習と評価には大量の計算が必要であるが、各世代においてそれぞれの個体は独立して学習・評価できることから、効率的な並列計算が可能である。

システム最適化にあたっては、例えば音声認識システムであれば認識誤り率を小さくすることは当然に重要であるが、同時に音声認識処理にかかる計算量やメモリ量なども削減したい場合がある。そのような場合は目的関数が複数となり、多目的最適化問題となる。単純には複数の目的関数を適当に重みづけして和を取ることで単一の目的関数に変換し単目的最適化問題として扱うことも考えられるが、目的関数の重み和を計算するための重みが新たなメタパラメタとなってしまう。そこで、新たな調整要素を導入することを避けつつ音声言語情報処理システムを最適化する仕組みとして、パレート最適と CMA-ES を組み合わせた手法を提案した。パレート最適を用いることで半順序集合において要素に順位付けを行うことができ、その順序を目的関数に用いることで最適化を行う。

CMA-ES およびパレート最適を用いて最適化した日本語音声認識システムは、成果発表の一環としてソフトウエアを一般公開している。同ソフトウエア（CSJ レシピ）は、日本語話し言葉コーパス(CSJ)のデータを用いてほぼ全自動でディープニューラルネットワークを用いた高性能日本語音声認識システムを構築し、認識評価実験を行うことができる。CSJ レシピは、音声認識研究者を中心として世界的な共同開発が行われている Kaldi ツールキットの一部として採用され、高性能な日本語音声認識システムとして国内外の研究期間や企業において、幅広く利用されている。

音声言語情報処理システムの半教師あり学習として、教師なし学習と教師あり学習を組み合わせた方法に取り組んだ。音声認識において古くからある半教師あり学習のフレームワークとして、教師あり学習した音声モデルを初期モデルとして用いて入力音声の認識を行い、得られた認識仮説を書き起こしラベルとして同じ入力音声を用いて教師あり学習を行う教師なし適応がある。ニューラルネットワークにおいてもこの方法を用いることができるが、最初の認識仮説の精度が良くない場合、教師なし適応によりかえって認識誤りが増えてしまう場合がある。そこで、出力層を分岐させたニューラルネットワークを用いたマルチタスク学習の枠組みを応用し、一つの出力層をラベルあり音声データに対する教師あり学習に、もう一つの出力層を認識仮説ラベルによる教師なし適応に用いる方法の検討を行った。同手法を用いることで単一の出力層を用いたのでは認識精度が劣化してしまう条件においても、認識精度を改善させられることを示した。

敵対的学習は、２つのニューラルネットワークをマッチポンプ式の競合状態に置くことで相互に競争させることで教師なし学習を実現する方法である。画像処理において大きな成功が示されている。同手法のニューラルネットワークによる任意話者声質変換への適用を検討し、課題はあるものの音声の明瞭性が改善すること示した。

音声認識システムにおける自動的な語彙獲得の実現を目的として、無限発音混合モデルを用いた発音辞書の半教師あり学習法の提案を行った。人間は言語学習の過程で対話などを通して最大で一日に 10 単語以上の単語を無意識に学習しているとされるが、同等の機能を音声認識システムにおいて実現することを目指したものである。提案法はノンパラメトリックベイズ法を応用したもので、推論にはギブスサンプリングを用いた。単語レベルのテキストデータと、それとは独立した音声データの音声認識結果の音素列をもとにした実験により、発音辞書中において発音が未知の単語の発音を自動的に補うことができることを示した。単語テキストと対応のある音素列から発音辞書を学習するのではなく、コンテキスト情報を手掛かりに対応の無い独立したデータから学習できる点が、本提案法の独自で新規な成果である。

言語リソースの比較的豊富な言語からリソースの乏しい言語への転移学習を目的として、DNN 音響特徴量の利用の検討を行った。日本語等のデータを用いて音韻性を強調し話者性を取り除く効果が期待されるように学習したニューラルネットワーク音響特徴量抽出器を、他の様々な言語のための音響特徴量抽出として応用した。ABX テストによる評価実験により、言語の違いにもかかわらず、従来の音響特徴量と比べて高い音素識別性能が得られることを示した。

既存の教師あり学習と比べて人手への依存度を減らすもう一つのアプローチとして、強化学習の検討を行った。提案法では、ネットワーク環境における多数の利用者を対象とした音声認識サービスを想定している。そこでは、利用者は起こしサービスを受けたい音声データをアップロードし、認識結果を受け取る。その際に、システムに対して意図的あるいは非意図的にごく簡単なフィードバックを行う。例えば認識の主結果と、もしそれが適切でなかった場合の次候補認識結果を比べて、適切と思う方を選択するなどである。利用者にとっては自分の欲しいテキストを選択しただけであるが、非常に多数の利用者からこのようなフィードバックを集めればシステムの学習に活用でき、開発コストの

かからない半自動のシステムが実現できると期待される。システムの学習法として方策勾配法を用いた手法を提案し、ユーザからのフィードバックをシミュレートした実験において有効性を示した。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 12 件）

① Taku Kato, Takahiro Shinozaki, “Reinforcement Learning of Speech Recognition System Based on Policy Gradient and Hypothesis Selection,” Proc. IEEE ICASSP, 査読有, pp. 5759-5763, 2018.

② Hayato Shibata, Taku Kato, Takahiro Shinozaki, Shinji Watanabe, “Composite Embedding Systems for Zerospeech2017 Track1,” Proc. IEEE ASRU, 査読有, pp. 747-753, 2017.

③ Hao Qin, Takahiro Shinozaki, Kevin Duh, “Evolution Strategy Based Automatic Tuning of Neural Machine Translation Systems,” Proc. International Workshop on Spoken Language Translation (IWSLT), 査読有, pp. 120-128, 2017.

④ Zhuang Bairong, Wang Wenbo, Li Zhiyu, Zheng Chonghui, Takahiro Shinozaki, “Comparative Analysis of Word Embedding Methods for DSTC6 End-to-End Conversation Modeling Track C,” Proc. Dialog System Technology Challenges (DSTC6), 査読有, pp. 1-5, 2017.

⑤ Takahiro Shinozaki, Shinji Watanabe, Daichi Mochihashi, Graham Neubig, “Semi-Supervised Learning of a Pronunciation Dictionary from Disjoint Phonemic Transcripts and Text,” Proc. Interspeech, 査読有, pp. 2546～2550, 2017.

⑥ Sou Miyamoto, Takashi Nose, Suzunosuke Ito, Harunori Koike, Yuya Chiba, Akinori Ito, Takahiro Shinozaki, “Voice Conversion from Arbitrary Speakers Based on Deep Neural Networks with Adversarial Learning,” Proc. Thirteenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 査読有, pp. 97-103, 2017.

⑦ 篠崎 隆宏, 渡部 晋治,「音声認識とブラックボックス最適化」, 日本音響学会誌 72 巻 10 号 pp. 644-652, 査読有, 2016.（解説論文）

⑧ Tomohiro Tanaka, Takahiro Shinozaki, Shinji Watanabe, Takaaki Hori, “Evolution Strategy Based Neural Network Optimization and LSTM Language Model for Robust Speech Recognition,” Proc. 4th International Workshop on Speech Processing in Everyday Environments CHiME 2016,” 査読有 pp.32-35, 2016.

⑨ Tomohiro Tanaka, Takafumi Moriya, Takahiro Shinozaki, Shinji Watanabe, Takaaki Hori, Kevin duh, “Automated Structure Discovery and Parameter Tuning of Neural Network Language Model Based on Evolution Strategy,” Proc. Spoken Language Technology (SLT) , 査読有, pp. 665-671, 2016. 10.1109/SLT.2016.7846334

⑩ Takafumi Moriya, Tomohiro Tanaka, Takahiro Shinozaki, Shinji Watanabe, Kevin Duh, “Automation of System Building for State-of-the-art Large Vocabulary Speech Recognition Using Evolution Strategy,” Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 査読有, pp. 610-616, 2015.

⑪ 齋藤 優貴, 能勢 隆, 伊藤 彰,「DNN を利用した Animation Unit の変換に基づく顔画像変換の検討」, 電子情報通信学会論文誌 D(レター), 査読有, Vol.J99-D, No.11 pp.1112-1115, DOI: 10.14923/transinfj.2016JDL8001

⑫ Takahiro Shinozaki, Shinji Watanabe, “Structure Discovery of Deep Neural Network Based on Evolutionary Algorithms,” 査読有, Proc. IEEE ICASSP, 2015.

〔学会発表〕（計 42 件）

① 加藤 拓, 篠崎 隆宏,「方策勾配法と仮説選択に基づく DNN 音声認識システムの強化学習」, 日本音響学会 2018 年春季研究発表会講演論文集, pp. 15-16, 2018.

② 鄭 崇輝, 李 知雨, 王 文博, 庄 佰 融, 篠崎 隆宏,「End-to-End ニューラル対話モデルにおける単語分散表現の比較検討」, 日本音響学会 2018 年春季研究発表会講演論文集, pp. 125-126, 2018

③ 宮本 颯, 能勢 隆, 千葉 祐弥, 伊藤 彰則,「DNN に基づく任意話者からの声質変換の雑音環境における評価」, 日本音響学会 2018 年春季研究発表会講演論文集, pp. 345-346, 2018.

④ 池下裕紀, 篠崎 隆宏, 渡部 晋治, 持橋 大地, Graham Neubig, 「ベイズ推論を用いた半教師あり学習の日本語適用」, 情報処理学会研究報告, Vol.2017-SLP-118, pp. 1-4, 2017.

⑤ 柴田 駿人, 加藤 拓, 篠崎 隆宏, 渡部 晋治,「ゼロリソース言語への応用を目的とした ABX テストによる DNN 特徴量の検討」,日本音響学会 2017 年秋季研究発表会

講演論文集，pp. 1-2，2017.
⑥ 覃 浩，篠崎 隆宏，Duh Kevin，「進化的戦略を用いたニューラル機械翻訳システムの自動最適化」，日本音響学会 2017 年秋季研究発表会講演論文集，pp. 1397-1398，2017.
⑦ 田中 智大，篠崎 隆宏，渡部 晋治，「Highway ネットワーク言語モデルを用いた日本語話し言葉音声認識」，日本音響学会 2017 年春季音響学会講演論文集，pp. 107-108，2017.
⑧ 小池 治憲，能勢 隆，伊藤 彰則，「読み上げ音声を用いたニューラルネットワークによる任意歌唱者歌声声質変換の検討」，日本音響学会 2017 年春季研究発表会講演論文集，pp. 357-358，2017.
⑨ 宮本 颯，能勢 隆，伊藤鈴乃介，小池治憲，伊藤彰則，「敵対的学習を利用したニューラルネットワークに基づく任意話者声質変換の検討」，日本音響学会 2017 年春季研究発表会講演論文集，pp. 355-356，2017.
⑩ 篠崎 隆宏，「音声認識ツールキット Kaldi を用いた大語彙日本語音声認識」，FIT2016，2016.（招待講演）
⑪ 田中 智大，森谷 崇史，篠崎 隆宏，渡部 晋治，堀 貴明，Kevin Duh，「進化的戦略を用いたリカレントニューラルネットワーク言語モデルの最適化」，日本音響学会 2016 年秋季音響学会講演論文集，pp. 31-32，2016.
⑫ 篠崎 隆宏，「大規模進化計算による音声認識システム開発の自動化」，GTC Japan 2016，2016.（招待講演）
⑬ 加藤 拓，篠崎 隆宏，「日本語話し言葉音声における半教師あり DNN 学習の検討」，情報処理学会研究報告，Vol. 2016-SLP-113，No. 1，pp.1-6，2016.
⑭ 篠崎 隆宏，「Kaldi ツールキットを用いた音声認識システムの構築」，電子情報通信学会/音響学会:音声研究会，2016.（招待講演）
⑮ 博多 屋涼，篠崎 隆宏，郡山 知樹，「粒子フィルタとガウス過程回帰によるシングルチャネル音源分離」，情報処理学会研究報告，Vol.2016-SLP-110，No. 6，2016.
⑯ 森谷 崇史，田中 智大，篠崎 隆宏，渡部 晋治，Duh，Kevin，「パレート最適と進化的戦略を用いた高精度大語彙音声認識システム構築の自動化」，電子情報通信学会/音響学会:音声研究会，SP2015-75，pp. 31-36，2015.
⑰ 齋藤 優貴，能勢 隆，篠崎 隆宏，伊藤彰則，「ビデオ通話における音声および表情特徴量を用いた話者変換の検討」，EMM 研究会，2015.
⑱ 小池 治憲，能勢 隆，篠崎 隆宏，伊藤彰則，「入力話者非依存ニューラルネットワークに基づく差分スペクトルフィルタを用いた声質変換における学習データ量の影響」，日本音響学会 2016 年春季研究発表会講演論文集，pp. 241-242，2016.
⑲ 齋藤 優貴，能勢 隆，篠崎 隆宏，伊藤彰則 「DNN を利用した Animation Unit の変換に基づく顔画像変換の検討」，電子情報通信学会技術研究報告，Vol. 115，no. 302，pp. 23-28 EMM 研究会，2015.
⑳ 伊藤 洋二郎，篠崎 隆宏，能勢 隆，「ニューラルネットワークを用いた話者特徴量抽出に基づく一対多クロスリンガル声質変換」，日本音響学会 2015 春季研究発表会講演論文集，pp. 397-398，2015.

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
www.ts.ip.titech.ac.jp

６．研究組織

(1)研究代表者
篠崎 隆宏 （SHINOZAKI，Takahiro）
東京工業大学・工学院・情報通信系・准教授
研究者番号：80447903

(2)研究分担者
能勢 隆（NOSE，Takashi）
東北大学・工学研究科・准教授
研究者番号： 90550591

(3)連携研究者
荒井 隆行（ARAI，Takayuki）
上智大学・理工学部・情報理工学科・教授
研究者番号： 80266072

(4)研究協力者
渡部 晋治（WATANABA, Shinji）
DUH, Kevin (DUH, Kevin)