様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　５　月　２３　日現在

機関番号：　１２６０８
研究種目：　基盤研究(B)（一般）
研究期間：　2013～2015
課題番号：　２５２８００５８

研究課題名（和文）Deep Generative Model とその因子分解による音声情報処理基盤

研究課題名（英文）Speech information processing using deep generative models and their factorization

研究代表者
篠田　浩一（Shinoda, Koichi）

東京工業大学・情報理工学（系）研究科・教授

研究者番号：１０３４３０９７

交付決定額（研究期間全体）：（直接経費）　　13,000,000円

研究成果の概要（和文）：多数話者の発声した大量の音声データから、Deep Neural Network (DNN)を構築し、それを音韻と話者の要因毎に分解することで高性能な音声認識モデルを獲得する枠組みの研究開発を行った。2つのDNNの一部を共有させた構造をもつDeep Siamese Networkを用いた話者認識、音韻構造を階層的な出力層で表現したDNNを用いた話者適応化、Soft Targetを教師としたStudent-Teacher学習の枠組みを用いた話者正規化学習、の3つの手法を提案し、それぞれで話者認識性能、音声認識性能の顕著な向上を得た。それ以外にもDNNの実装、ネットワーク構造設計の研究を行った。

研究成果の概要（英文）：In speech recognition, it is important to train an accurate deep neural network (DNN) acoustic model from a large amount speech data from many speakers. In this study, we developed a framework to improve accuracy of the DNN acoustic model by factorizing speech data into phoneme and speaker elements. First we developed a speaker recognition method using deep Siamese network in which two DNNs which share its part. Second, we applied a DNN with a hierarchical phonetic structure to speaker adaptation. Third, we developed a speaker-adaptive training method where we utilized a student-teacher learning framework using soft targets. We improved speaker verification and speech recognition performance. We also studied DNN implementation and DNN structure design.

研究分野：　音声情報処理

キーワード：　音声情報処理　深層学習　話者適応

１．研究開始当初の背景

研究開始当初(2013 年春)、音声認識において、深層学習(Deep Learning)が顕著に性能向上に寄与することが明らかになり、多くの研究機関がその研究に本格的に取り組み始めた。例えば、Google, Microsoft, IBM など世界の有力研究機関がその例である。Deep Learning が音声研究にパラダイムシフトをもたらす可能性が指摘されていた（それは 3 年後の今、現実となっている。）。

我々はその 2 年前から研究を開始しており、日本語音声の認識においてその効果を既に確認していた。その理論的背景は明らかでない。本研究は、他に先駆けて深層学習による新しい方法論とそれに基づく技術基盤を確立したい、という動機のもと開始された。

２．研究の目的

多数話者の発声した大量の音声データから、Deep Neural Network (DNN)を構築し、それを音韻と話者の要因毎に分解することで高性能な音声認識モデル・話者認識モデルを獲得する枠組みの研究開発を行う。

３．研究の方法

(1) Deep Siamese Network から抽出した話者情報を用いた話者同定

音声から同じ話者が発声した音声かどうかを判断する話者同定技術は、個人認証の 1 つの手段として有望である。ここでは多数話者の音声から学習された DNN を用いて、音声特徴を音韻特徴と話者特徴に分離し、話者特徴のみを取り出し、それを話者同定のための特徴量として用いる手法を検討した。

従来の話者同定では、混合ガウスモデル（Gaussian Mixture Model; GMM)とサポートベクターマシン(Support Vector Machine; SVM)を組み合わせた GMM-SVM システムがしばしば用いられる。入力としては、通常、メル周波数ケプストラム係数(Mel-Frequency Cepstral Coefficients; MFCC)が用いられる。

MFCC は、一般の音声特徴を表現する特徴量であり、それには、音韻特徴や話者特徴、マイク（回線）の違いを表す特徴、周囲雑音に関する特徴など、様々なものが含まれていると考えられる。この中で、音韻(phoneme)とは、音声をテキストに書き起こしたもの（ここでは言語と呼ぶ）を構成する最小の音声単位であり、音韻特徴は音声に含まれる言語情報を表現する特徴である。また、話者特徴とは、音声に含まれる話者に関する情報を表現する。ここでは、マイク（回線）や周囲雑音など他の要因は一定であるという条件下で、MFCC から音韻特徴と話者特徴を分離して話者特徴を取り出す。

そのために、多層ニューラルネットワーク(Deep Neural Network; DNN) の一種である Regularized Deep Siamese Network (RDSN) を用いる。このネットワークは、2 つの多層 Denoising AutoEncoder (DAE)から構成され、その中間層の一つの層において、一部のノードが 2 つの DAE の間で共有される構造をもつ。ここで DAE とは、ノイズが重畳した信号を入力とし、そこからノイズを取り除いた(クリーンな)信号を出力するネットワークのことである。

このネットワークは以下のように学習される。まず、1 つの多層 DAE の学習を行う。ここでは、クリーンな発声と、それに雑音を人工的に重畳したノイズ重畳発声のペアを多数用意し、後者を入力、前者を出力として用いて、重み係数の推定を行う。次に、構成した DAE をコピーしてもう一つ作り、その中間層のノードの一部を共有する RDSN を構築する。ここで共有したノード群を共有ノードと呼ぶ RDSN の重み係数の推定では、同一の話者の発声ペアと違う話者の発声ペアを多数用いて行われる。ここで、ペアを構成する 2 つの発声の言語情報は一般に異なる。ペアの一方が RDSN を構成する片方の DAE に入力され、もう一方がもう片方の DAE に入力される。

学習の目的関数として以下の 2 つの目的関数 LR と LC を重み付したものを用いる。LR は、各々の DAE における入力と出力との間の差分を全発声について足し合わせたものである。LC は、発声ペアの各々の発声に対する共有ノードの出力について、同一話者のときはその差分、違う話者のときには差分の増加に対して単調減少する関数の出力を、すべての発声ペアについて足し合わせたものである。この目的関数を最小化するように学習を行う。つまり、共有ノードの出力が各々の話者固有の特徴を表すものとなり、2 つの DAE の出力が話者によらない音韻特徴を表すものになるように重み係数を学習する。

このようにして学習された RDSN の共有ノードからの出力を、MFCC の代わりに GMM-SVM システムの入力として用いることで、話者同定の性能が向上することが期待される。

(2)階層的な出力層をもつ DNN を用いた話者適応

使用者の少量の発声を用いて音声認識の性能を向上させる話者適応技術は、深層学習が盛んになる以前から盛んに研究され、実用化されてきた。深層学習を用いた音声認識においても、話者により性能が異なることが報告されており、話者適応技術の重要性は高い。

深層学習においては一般に推定すべき重みパラメータ数が多く、少量の話者の発声で学習しようとすると、過学習という現象が起き、性能が却って劣化してしまう。過学習を起こさないためには推定すべきパラメータ数を絞る必要がある。

従来法の多くは、DNN の入力層のすぐ上の層の出力を入力に対して線形にし、その 2 層

の間の重み係数のみを使用者の発声から推定する方法である。これは、従来用いられている特徴量空間最ゆう線形回帰(feature-space Maximum Likelihood Linear Regression; fMLLR)と同様、話者間の変換を特徴量空間における線形変換に近似し、その変換行列を求めることに相当する。つまり、話者要因の変動をDNNの下層における線形変換で吸収している。この種の手法は発声から話者特徴を除去するものであり、話者正規化と呼ばれる。推定パラメータ数が全体の学習に比べはるかに少ないため安定に学習できる。しかし、話者間の変換は実際には線形変換とは異なるための、その性能向上の度合いに限界がある。

一方、音声からの音韻特徴と話者特徴の分離という観点からみると、音声認識性能の向上のため数多くの音韻モデルが個別にモデル化されており、少ない使用者の発声で、それらの数多くの音韻モデル各々における話者による特徴の変化を十分に学習できないことが過学習の原因と考えられる。現在主流の音声認識では、音韻モデルとして環境依存音素モデルが用いられている。このモデルでは音素をモデル化する際に、その音素のみではなく、その前後にある音素も考慮してモデル化されている。つまり同じ音素"a"でも、その前後にある音素が違ったら別々にモデル化されている。この場合モデル数は通常3000～6000個となる。話者の少量の発声（例えば10文程度)では、全部のモデルに対する発声が十分あることはなく、通常は、ほんの少ししか学習データがないモデルや、まったく学習データがないモデルが存在する。これが性能低下を引き起こす原因である。

研究代表者らは先に隠れマルコフモデルを用いた音声認識において、Structural Maximum A Posteriori (SMAP)推定を用いた話者適応法を提案している。この方法では、まず、音韻の構造を木構造の形で明示的に表現する。例えば、(環境に依存しない)音素を親ノード、同じ中心音素をもつ環境異存音素を子ノードとする木構造を構築する。その上で、親ノードに対応するパラメータを事前パラメータとして、その子ノードのパラメータを、その事後確率が最大になるように求める。このように、より多くのデータを用いて推定されたパラメータを利用して、より少ないデータしかないパラメータを推定することにより、データ量が少ない場合でも頑健な推定が実現される。

提案手法では、この SMAP 話者適応の考え方を深層学習に応用する。まず、予め多数話者の大量の発声データを用いて、音声認識のためのDNNを学習する。このDNNの出力層の各ノードは、環境依存音素の各状態に対応する。そして、その出力層の上に、音素の各状態に対応するノードからなる新たな出力層を置く。適応時には、新たに追加した出力層における教師ラベルを用いて、少量の話者の発声を用いたDNNの学習を行う。認識時には、新たに追加した出力層は取り去り、元の DNN の環境依存音素の状態に対応する出力ノードからの出力を用いて音声認識を行う。このように音素の制約をかけて環境依存音素モデルを学習することにより、少ないデータで学習する場合でも過学習に陥らない頑健な学習が可能になる。

(3)DNNに対するStudent-Teacher学習を用いた話者正規化

DNN を用いて話者の違いに対して頑健な音声認識を実用化しようとする場合、前にある程度の量の話者の発声をいったん集める必要がある。また、音声認識のためのDNNは一般に規模が大きく、そのパラメータ学習に時間がかかる。オンラインで高速に話者の発声に適応する手法に対する需要は大きい。ここでは、まず、soft-target 学習を用いて DNN の規模を認識性能の劣化を抑えつつ小さくする。次に、話者正規化後の特徴量を入力として学習されたDNNからの出力を教師として、話者正規化をしない特徴量を入力とした DNN を学習する転移学習を行う。これらの2つの手法により、計算量が小さく、かつ、オンラインで動作する認識が可能となる。以下、各々について説明する。

まずsoft-target学習では、まず大量の音声とそれに人手で付与された教師ラベルを用いて大規模な高性能DNNを構築する。その上で、同じ音声を入力として用い、それに対する大規模DNNの出力を教師信号として、より小さいサイズのDNNを学習する。この小規模DNNは、入力層と出力層のノード数は大規模DNNと同じであるが、隠れ層の数や、各隠れ層のノード数が大規模 DNN に比べ小さい。この人手で付与されたラベルを教師として用いる場合、出力層の各ノードに対応する教師信号は0または1となるが、大規模DNNの出力を教師として用いる場合は、教師信号は0と1の間の実数となる。そのため、この学習はsoft-target学習と呼ばれる。これに対し、0,1 の教師信号を用いる場合はhard-target学習と呼ばれる。

Soft target は hard target に比べ、より多くの情報を含み、従って、小規模DNNの学習にはsoft-target学習のほうがより効果があることが期待できる。さらに、soft-target学習では、教師ラベルの付いていないデータを活用できるという利点もある。つまり、一旦教師ラベル付きデータで学習した大規模DNN が構築できれば、各音声に対する小規模DNN の学習のための教師信号を得るために人手により付与された教師ラベルは必要ない。一般に教師ラベルのついていない音声データは教師ラベルのついていない音声データに比べ豊富にあるため、それらを用いることにより、一層の性能向上が望める。

このようにSoft target学習で構成された小規模DNNは計算量が少なくオンラインでの

動作が可能である。これに加えて話者正規化処理も同時に行う学習を行う。まず、特徴量空間最ゆう線形回帰(fMLLR)により話者正規化された音声を用いて、大規模DNNを学習する。次にSoft target学習を行う際、入力として話者正規化された音声を用いる代わりに話者正規化前の(話者性を含む)音声を用いる。ここで、教師信号として用いる大規模DNNの出力は、話者正規化された音声を用いて得られたものを用いる。この処理で、話者正規化の処理をも行う小規模DNNを構築することが可能になる。

(4)音声センサ

センサーネットワークのノードとして音声認識センサを用いる場合、様々な雑音が存在する環境下で稀に入力されるコマンド音声を認識することや、限られたハードウエア上で極めて少ない消費電力で動作することなど、一般的な音声認識とは大きく異なった条件での認識処理が求められる。そのような条件で精度の高い認識性能を実現する認識アーキテクチャを実現した。

DNNを用いた新しいキーワードスポッターを提案するとともに、その実証のためハードウエア実装を行いた認識実験を行った。将来的にはMEMSマイクロホンと一体化した音声認識センサの実現を想定するが、実験では様々な認識アーキテクチャを試すためMEMSマイクロホンとFPGAを組み合わせた回路を実装して用いた。

(5) 神経核ネットワークを模したニューラルネットによる音声認識

音声認識において用いられるニューラルネットワークは、複数のニューロンレイヤーが一列に並んだ構造をしたものにほぼ限られている。そこでより柔軟な構造を用いることで、認識性能の向上や計算量の削減を図った。

構造の柔軟性とGPU上での計算効率を両立させるネットワーク構造として、ニューロン配列をノードとした有向無サイクルグラフにより表現される神経核ネットワーク型フィードフォワードニューラルネットワークを提案し、音声認識に応用することでその有効性について評価した。

４．研究成果

(1) Deep Siamese Networkから抽出した話者情報を用いた話者同定

RDSNから抽出した話者特徴の効果を調べるために、評価実験を行った。データとしては話者同定のベンチマークでよく用いられるNIST-SREデータセットの2004年版と2006年版を用いた。評価話者として100名の男性話者の発声を用いた。MFCCを用いたGMM-SVMに比べ、提案手法は5%誤りを減少させた。さらに、MFCCを用いたシステムと提案システムの結果を融合して用いることにより、誤りを6.6%削減した。

この枠組みは音韻特徴と話者特徴の分離以外にも様々な用途に適用でいる。例えば、提案手法では、周囲雑音やマイクの違いを考慮していないが、これらも同様の手法で抽出・分離することにより、より高性能な話者特徴を取り出せる可能性がある。また、現在は10msの音声フレーム毎の入力を用いているが、多数のフレームを同時に入力することにより、更なる性能向上が期待できる。

(2) 階層的な出力層をもつDNMを用いた話者適応

提案手法を音声サーチと音声対話の500時間のデータを初期DNNの学習に用い、それに含まれない59名の話者の発声した40文を適応に、125文を認識評価に用いて評価を行った。初期DNNに対し適応データをそのまま用いて
学習した場合と比べ、顕著な性能向上を得た。

今後は、更に効果的な音韻構造の利用を検討する必要がある。また、前述の、
DNNの下層に線形変換を置き、そのパラメータを推定する手法との組み合わせも有望である。

(3) DNNに対するStudent-Teacher学習を用いた話者正規化

提案手法を英語発声の大規模音声データセットを用いて評価した。話者正規化を行わない、hard-target学習を行った大規模DNNと比較して、提案手法は5.3%のエラーを削減した。また、同時にパラメータ数は12.2%となり、高性能化と小規模化を同時に達成した。

今後は、系列学習の導入や、多くの違う構造をもつ小規模DNNを組み合わせることなどにより、更なる性能向上が期待できる。

(4) 音声センサ

DNNを用いた単語検出器を提案し、ハードウエア記述言語を用いて設計・実装した。別途用意したDNNを用いた認識実験により、提案認識器により従来法よりも優れた認識精度が得られることを示した。

(5) 神経核ネットワークを模したニューラルネットによる音声認識

ネットワーク構造を適切に設計できれば、全体としてより少ないニューロンユニットを用いたネットワークでより高い認識精度が得られることを示した。

５．主な発表論文等

〔雑誌論文〕(計　4件)

① R. Price, K. Iso, K. Shinoda, "Wise teachers train better DNN acoustic models", EURASIP Journal on Audio Speech, 査読有, p. 1-19, 2016.

DOI:10.1186/s13636-016-0088-7

② T. Shinozaki, S. Watanabe, "Structure discovery of deep neural network based on evolutionary algorithms", Proc. ICASSP, 査読有, pp.4979-4983, 2015. 10.1109/ICASSP.2015.7178918

③ R. Price, K. Iso, K. Shinoda, "Speaker adaptation of deep neural network using a hierarchy of output layers", Proc. IEEE Spoken Language Technology Workshop, 査読有, pp. 153 - 158, 2014. 10.1109/SLT.2014.7078566

④ R. Price, S. Biswas, K. Shinoda, "Combining deep speaker specific representation with GMM-SVM for speaker verification", Proc. INTERSPEECH2013, 査読有, pp. 2788-2792, 2013.

〔学会発表〕(計 11 件)

① 森谷崇史, 田中智大, 篠崎隆宏, 渡部晋治, Duh Kevin, "進化的戦略による高精度大語彙音声認識システムの多目的最適化", 日本音響学会 2016 年春季研究発表会,2016 年 3 月 9 日, 桐蔭横浜大学 (神奈川横浜市)

② 田中智大, 森谷崇史, 篠崎隆宏, 渡部晋治,堀貴明, "Kaldi 用 CSJ レシピへの RNN 言語モデルの導入と性能評価", 2016 年春季研究発表会, 2016 年 3 月 9 日, 桐蔭横浜大学 (神奈川横浜市)

③ 篠田浩一,"(招待講演)音声・画像・映像における Deep Learning を用いたパターン認識", 人工知能学会 AI チャレンジ研究会, 2015 年 11 月 12 日, 慶応大学(神奈川県・横浜市).

④ D. Hoesen, R. Price, R. L. Dessi, K. Shinoda, "A DNN-based ASR system for the Indonesian language", 日本音響学会 2015 年秋季研究発表会, 2015 年 9 月 16 日, 会津大学 (福島県会津若松市).

⑤ 松山祐輔, R. Price, 篠田浩一, "活性化関数のパラメータ制御を用いた LSTM による音声認識",日本音響学会 2015 年秋季研究発表会, 2015 年 9 月 16 日, 会津大学 (福島県会津若松市).

⑥ 朱凱, 李昊霖, 篠崎隆宏, 堀内靖雄, 黒岩眞吾, "DNN 特徴量抽出器に基づく単語検出器の FPGA 実装と評価", 日本音響学会 2015 年 9 月 16 日, 会津大学 (福島県会津若松市).

⑦ 篠田浩一,"(招待講演)音声認識のための Deep Learning", 第 25 回日本神経回路学会全国大会, 2015 年 9 月 4 日, 電気通信大学 (東京都・府中市).

⑧ 福田峻, 井上中順, 篠田浩一, "CNN から抽出した複数特徴量の統合に基づいた映像の意味インデクシング", 第 21 回画像センシングシンポジウム(SSII), 2015 年 6 月 11 日, パシフィコ横浜アネックスホール (神奈川県・横浜市).

⑨ 篠田浩一, "(招待講演)統計的パターン認識のための中間表現", 電子情報通信学会 2015 年 3 月 SIP/AE/SP 研究会, 2015 年 3 月 2 日、石垣島ホテルミヤハラ (沖縄県・石垣市).

⑩ 篠田浩一, "(招待講演)Deep Learning による新しい音声認識パラダイム)", 日本神経回路学会主催セミナー「Deep Learning が拓く世界」, 2014 年 8 月 26 日、京都大学東京オフィス (東京都・品川区).

〔その他〕
ホームページ等
http://www.ks.cs.titech.ac.jp/japanese/index.html

６．研究組織

(1)研究代表者
篠田　浩一 (SHINODA, Koichi)
東京工業大学・大学院情報理工学研究科・教授
研究者番号：10343097

(2)研究分担者
岩野　公司 (IWANO, Koji)
東京都市大学・メディア学部・教授
研究者番号： 90323823

篠崎　隆宏 (SHINOZAKI, Takahiro)
東京工業大学・大学院総合理工学研究科・准教授
研究者番号： 80447903