博士学位論文審査要旨

2018年 1月 23日

論文題目： Deep Learning Algorithms for High-performance Automatic Speech Recognition
（高性能自動音声認識のための深層学習アルゴリズム）

学位申請者： 落合 翼

審査委員：

主 査：同志社大学大学院理工学研究科 教授 片桐 滋

副 査：同志社大学大学院理工学研究科 教授 加藤 恒夫

副 査：国立研究開発法人情報通信研究機構先端的音声翻訳研究開発推進センター 主任研究員 Lu Xugang

要 旨：

人間とコンピュータとの自然な対話の実現を目指す音声認識技術の研究においては，長年，隠れマルコフモデル（HMM: Hidden Markov Model）を用いて音声認識システムが構築されてきた．こうした状況の中で近年，豊かな特徴表現力で知られる深層ニューラルネットワーク（DNN: Deep Neural Network）が登場し，それを用いた様々な音声認識の試みが行われている．しかし，そうしたDNNを用いる音声認識システムでさえも，発話者や発話環境などに因る入力音声の変動にはまだ十分に対処できていない．

本論文は，その不十分さの解決を目指す２つのアプローチ，即ち１）話者適応学習（SAT: Speaker Adaptive Training）法に基づくハイブリッドDNN-HMM音声認識（SAT-DNN-HMM）法と，２）雑音抑制ビームフォーマとDNN型音声認識部を認識性能向上の観点で統合的に学習するマルチチャンネル・エンドツーエンド音声認識（ME2E: Multi-channel End-To-End）法とを提案し，それぞれの有効性を実証するものである．

SAT-DNN-HMM法は，ほぼ一様な内部構造を持つ従来のDNNの内部に，発話者毎に異なる話者モジュールを局在化させ，その話者モジュールが話者適応段階で入れ替えられることを前提とした学習を行う手法である．およそ340名の発話者データを用いた体系的な話者適応実験を通して，本手法の有効性を実証している．

なお，本手法における話者モジュールは，元々のDNNの一部である非線形ネットワークを再利用している．本論文では，その非線形ネットワークに代えて新たに挿入する線形変換ネットワークを話者モジュールとして利用する手法も提案している．上記と同様の実験を通して，この線形変換ネットワークの挿入がSAT-DNN-HMM法の性能を一層向上できることを明らかにしている．

大きな特徴表現力を確保するため，必要以上に大きなDNNが用いられることが多い．しかし，その過大なDNNの利用は，しばしば学習や適応を困難にする．この問題に着目し，本論文では，ネットワーク重み行列に関する特異値分解による次元圧縮によって，SAT-DNN-HMM法のDNNをその1/4程度にまで大幅に圧縮し，その結果として削減前よりも効率的に話者適応が可能であることも明らかにしている．

ME2E法は，ニューラル・ビームフォーマとモジュール内包型のDNNとを組み合わせ，かつ入力音声チャンネルとビームフォーマの出力である雑音抑制音声チャンネルを並存させた上で認識性能向上を目指す統合学習を行う手法である．鉄道駅舎などの雑音を避け得ない様々な公共空間において収録した音声を用いた実験を通して，本提案手法が，先端的雑音抑制システムを前段にもつ音声認識システムと同様の高い音声認識精度を，エンドツーエンドの枠組みにおいて自

動的に達成し得ることを明らかにしている.

また，本手法が認識精度の向上を目指した学習を行っているのみにもかかわらず，同時に雑音を抑制したクリーンな音声の生成をも可能としていることも明らかにしている.

以上の成果は，音声認識あるいは機械学習，人工知能の技術の発展に多大なる貢献を為すものである．よって本論文は博士（工学）（同志社大学）の学位論文として十分な価値を有するものと認められる.

総合試験結果の要旨

2018年 1月 23日

論 文 題 目： Deep Learning Algorithms for High-performance Automatic Speech Recognition
（高性能自動音声認識のための深層学習アルゴリズム）

学位申請者： 落合 翼

審 査 委 員：

主 査：同志社大学大学院理工学研究科 教授 片桐 滋

副 査：同志社大学大学院理工学研究科 教授 加藤 恒夫

副 査：国立研究開発法人情報通信研究機構先端的音声翻訳研究開発推進センター
主任研究員 Lu Xugang

要 旨：

本論文提出者は，理工学研究科博士前期課程を修了している．本論文の主たる内容は，IEICE Transactions on Information and Systems や IEEE Journal of Selected Topics in Signal Processing，Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 等の当該分野において中心的な論文誌・会議録に掲載され十分な評価を受けている．2018年1月13日午前10時から2時間にわたり学術講演会が開かれ，種々の質疑討論が行われたが，提出者の説明により，十分な理解が得られた．講演会終了後，審査委員により学位論文に関連した諸問題につき口頭試問を実施した結果，十分な学力を確認できた．提出者は，英語による論文発表や語学試験の合格も果たしており，十分な語学能力を有すると認められる．よって，総合試験の結果は合格であると認める．

博 士 学 位 論 文 要 旨

論 文 題 目： Deep Learning Algorithms for High-performance Automatic Speech Recognition
(高性能自動音声認識のための深層学習アルゴリズム)

氏　　名： 落合　翼

要　　旨：

Speech is the most natural communication modality for humans. Since the advent of computers, the development of natural speech communication channels between humans and computers has been one of the greatest goals in computer science research fields. For such natural communication channels, automatic speech recognition (ASR) technology, which converts speech into text by computer programs, has been scrutinized and comprises the core of such intelligent and user-friendly applications as voice dictation, voice search, voice command, and spoken dialogue systems.

Over the past few decades, machine learning (ML)-based approaches have been a center pillar of ASR research, based on the advancement of such key ML methodologies as (artificial) neural networks and probabilistic modeling. With the recent advent of deep learning (DL) techniques, ASR performances have significantly improved in the past five years or so. However, such DL-based ASR technologies remain insufficient for appropriately coping with the variability of speakers and speaking environments; ASR technologies must be improved.

The most fundamental way for coping with the variability is to fully represent it in training stages for ASR systems. However, it is basically unrealistic to predict the entire variability. Accordingly, incorporating some effective compensation techniques with DL-based ASR systems is a reasonable solution to the variability problem. Motivated by this understanding, in this dissertation, we investigate novel compensation techniques for deep neural network (DNN)-based ASR systems by introducing an internal structure to DNN, which enables ASR systems to evoke the aptitude of DNN for dealing with speaker and/or environment variability.

In this dissertation, we separately study two kinds of variability: 1) variability in speakers, and 2) variability in speaking environments (changes in background noise and reverberation). For the former issue, we propose a novel speaker adaptation algorithm that incorporates the speaker adaptive training (SAT) concept into the training of DNN in the framework of a hybrid DNN and Hidden Markov Model (HMM) speech recognizer. Our proposed SAT-based speaker adaptation scheme introduces modularity, more precisely, localizing a speaker dependent (SD) module, in the DNN part of the hybrid system and optimizes the DNN part, assuming that the SD module is adapted in the adaptation stage. For the latter environment-related issue, we focus on a recently proposed end-to-end ASR architecture, which is completely composed of neural networks, and propose a novel

multichannel end-to-end (ME2E) ASR architecture that integrates speech enhancement and speech recognition components into a single neural network-based architecture. Our proposed architecture allows the overall procedure of multichannel speech recognition (i.e., from speech enhancement to speech recognition) to be optimized only under a recognition-oriented training (learning) criterion using multichannel noisy speech samples and corresponding transcriptions.

We conducted several experimental evaluations of our proposed methods and successfully demonstrated their effectiveness in further increasing the performances of DNN-based ASR systems. The proposed SAT-based speaker adaptation methods successfully increased the adaptability of hybrid DNN-HMM ASR systems and reduced the size of the adaptable parameters. The proposed ME2E ASR architecture successfully learned a noise suppression function through end-to-end recognition-oriented optimization and improved ASR performances in various noisy environments.