コンテキスト適応型ニューラルネットワークを用いた音声インタフェースのパーソナライズ化

近年，音声インタフェースの普及が急速に進んでいますが，音声認識結果の詳細を分析すると，必ずしもすべての話者に対して同程度の音声認識性能が得られているわけではなく，うまく認識ができる話者，できない話者がいるのが現実です．この問題に対応するため，NTTコミュニケーション科学基礎研究所ではニューラルネットワークの新たな構造に関する研究を行い，コンテキスト適応型ニューラルネットワークを開発しました．本稿では，自動音声認識システムの中で用いられる音のモデル（音響モデル）の話者適応に関する最新研究成果を紹介します．

Marc Delcroix　/木下（きのした） 慶介（けいすけ）
小川（おがわ） 厚徳（あつのり）　/苅田（かりた） 成樹（しげき）
樋口（ひぐち） 卓哉（たくや）　/中谷（なかたに） 智広（ともひろ）
NTTコミュニケーション科学基礎研究所

音声認識システム

近年，音声認識システムは急速に普及しており，私たちの日常生活の中でも，それらが活用されている場面が増えてきました．例えば，天気予報や近くのお勧めレストラン情報を知りたいときに，スマートフォンに話しかけ，質問を投げかけるのはもはや当たり前の光景になってきています．また，ホームアシスタントや音声対話ロボットなどのコミュニケーションエージェントの普及も進んできています．私たちが何かしらの情報にアクセスするための主な手段が「音声」になっていく未来はそう遠くはないのかもしれません．

昨今の深層ニューラルネットワーク（DNN:Deep Neural Network）技術[1]の発展により，音声認識の精度は飛躍的に向上し，それに伴い音声認識を用いた製品の開発や普及も急速に進んできました．しかしながら，DNNを用いることで，すべての問題が解決したわけではありません．現状においても，音声認識性能が，話者の声質や周囲の騒音環境などの音響コンテキストによって大きく変動してしまうという問題は，まだまだ解決されていません．

本稿では，音声認識システムをそれらの音響コンテキストに即座に適応させるためにNTTコミュニケーション科学基礎研究所で開発したコンテキスト適応型DNN（CADNN）[2]を紹介します．CADNNは，ニューラルネットワークのパラメータを，話者の声質特徴や周囲の騒音特徴といった外部のコンテキスト情報に応じて変更させることのできる新たなニューラルネットワークです．これによって，発話者の音声を認識するために最適な音声認識システムを即座につくり出すことが可能となり，さまざまな環境において高い音声認識性能を確保することができるようになります．

DNNを用いた音響モデル

■構成要素

音声認識システムは，いくつかの要素で構成されています（図1）．1番目の構成要素は特徴量抽出モジュールです．このモジュールは，音声発話を30 ms程度の短い時間フレームに区切り，それぞれのフレームから音声特徴量を抽出します．2番目の構成要素は音響モデルです．音響モデルは，入力の各時間フレームの音声特徴量がどの音素に対応しているかを計算し，その情報を確率値として出力します．3番目の構成要素はデコーダです．デコーダは入力された特徴量系列を基に，最適な単語系列を発見する役割を担っており，具体的には音響モデルが出力した音素の確率，音素系列と単語の関係を表した発音辞書，ある単語とある単語がつながる確率を出力する言語モデルからの情報を基に，もっとも適切と思われる認識結果を出力します．次に，

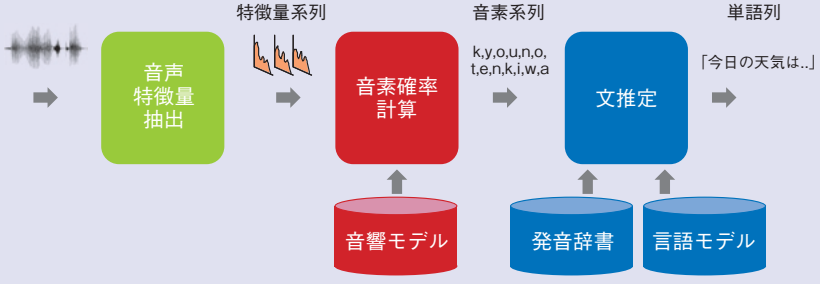


図1　音声認識システム

本稿の主題である音響モデルに焦点を当て，話者適応を題材に議論を進めます．

■音響モデルの例

近年の音響モデルの中では，入力の音声特徴量を音素の確率に変換するためのマッピングツールとしてDNNが使われています．DNN型の音響モデルの一例を**図2**に示します．多くの場合，DNNはいくつかの隠れ層から成り，その複数の隠れ層が入力信号を非線形に変換する働きを担います．隠れ層を多重にすれば，入力信号と出力信号（所望信号）の間の複雑な（非線形な）関係を正確にとらえることが可能となります．

音響モデルにおいては，DNNの入力とは音声の特徴量であり，出力は音素の確率です．数十時間から数千時間に及ぶ音声データと各音声の発話内容の書き起こし，さらにはその書き起こし情報から得られる，実際に話された音素系列の情報を用いて，DNN型音響モデルの最適化（学習）は行われます．学習には，確率的勾配降下法（SGD:Stochastic Gradient Descent）を用いた誤差逆伝搬法を用いるのが一般的です．

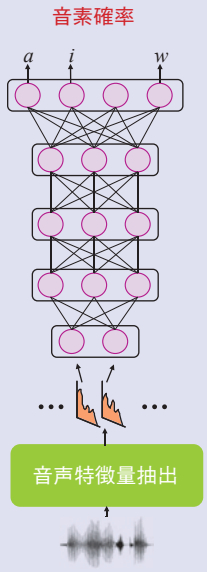


図2　DNNに基づく音響モデル

■音響モデル構築の課題

さまざまな話者の声を正確に認識できる音響モデルを構築するためには，一般的には，たくさんの話者の声を含むような学習データを用意する必要があります．そして，そのような学習データを用いれば，平均的には良い振る舞いをする音響モデルを構築することができます．しかし，そのような学習方法では，音声認識システムを今まさに使おうとしているある特定のユーザの声にシステムが最適化されている保証はなく，結果としてはうまく認識ができる話者，できない話者が出てきてしまうのが常です．この問題を解決するためには，音響モデルをユーザの声に合わせ込む（適応させる）必要があります．しかしながら，各ユーザに事前に何時間もの音声を発話してもらい，かつその発話内容の書き起こしデータを用意するようなことは一般的には困難であり，結果，そのようなかたちでユーザの声に合わせた音響モデルをつくることも同様に困難です．実際に多くのアプリケーションでは，話者適応は，数秒の音声データを基に行わなければならない場面が多く，またその数秒のデータの正確な書き起こし情報も手に入れることはできません．高速な教師なし話者適応技術の誕生が待ち望まれています．

CADNN

■CADNNの基本的な考え方

音響モデルの話者適応に関する研究は，さまざまな研究機関にて行われています．近年提案されている方法の中で有望と思われるアプローチは，ユーザの声の特徴を表す補助的特徴量をDNNの入力に付け加え，DNNにそのユーザの声質に関する情報を明示的に教える方法です．このような話者の声質に関する情報は，わずか数秒の音声データから抽出することができ，また抽出のためには発話内容の書き起こしは必要でないところが，このアプローチの利点です．しかし，このような補助的特徴量をDNNの入力に付け加えるだけでは，DNNの中のパラメータのほんの一部にしか影響を与えることができず，効果は限定的です．私たちは，新たな枠組みであるCADNNを開発し，この補助的特徴量を別のかたちで活用する方法を考案しました．

CADNNの基本的な考え方は，ある音響コンテキスト（例えば，話者）に合わせて学習したDNNは，その音響コンテキストの音声を認識するためには最適である，というものです．この考え方自体は一般的なもので，例えば，男性の声の認識に特化したDNNを男性の声のみを用いて作成し，女性の声に特化したDNNは女性の声のみを用いて作成すれば，2つの異なるコンテキストに最適と思われるDNNをつくることができます．このような場合，ユーザの性別に合わせて，使用するDNNを選択すれば，ある種の話者適応を実現することができます．しかし，このような単純な手法には2つの問題点があります．まず，このように別々のモデルを作成する方策を取ると，各モデルの学習に使用できるデータは，全データのうちのほんの一部のみとなってしまいます．一般的には，男性

の声と女性の声には異なる部分もありますが，共通な部分も多く，その共通部分を学習するためには全データを使用したほうが良いことは明白です．しかし，別々のモデルを作成する方策を取ると，それができなくなるのです．また，2つではなく，もっと多くのDNNを用意した場合，入力の音声をもっともうまく認識できるであろうDNNをその中から選び出すこと自体も非常に困難な作業となります．

■CADNNの仕組み

CADNNは，音響コンテキストに依存する部分を，ネットワークの一部のみに限定することで，前述の1番目の問題に対処します．また，音響コンテキストに依存する部分の中のどの部分をより重視して計算を行うかを決定するのは，補助的特徴量を入力に取る別のニューラルネットワークです．提案CADNNの概略を**図3**に示します[(3)]．CADNNの特徴は，分解された隠れ層（サブ隠れ層）を持っている点です．各サブ隠れ層は，ある音響コンテキストに紐付いています．例えば，性別を音響コンテキストとした場合，考慮したい音響コンテキストの数は2となるため，男性用に1つのサブ隠れ層を，女性用にもう1つのサブ隠れ層を用意します．各サブ隠れ層からの出力結果は，補助的特徴量から計算される重み係数を用いて足し合わされ，この隠れ層からの最終的な出力となります．この重み係数の計算は，補助的特徴量を入力とする小さな補助ネットワークが行います．この重み係数は，音声認識性能がもっとも高くなるように計算されるよう，補助ネットワークとCADNNは学習時には一体的に最適化されます．

CADNNにはいくつかの長所があります．1番目の長所は，補助ネットワークとCADNNは重み係数を通じてつながっているため，学習時には同時に最適化することができる点です．つまり，与えられた入力音声と，そこから抽出された音響コンテキストを表す補助的特徴量をCADNNと補助ネットワークにそれぞれ与えれば，音声認識精度が最大となるように，CADNNのパラメータ，音響コンテキストの重み係数計算のための補助ネットワークのパラメータすべてを一挙に最適化することが可能です．また，2番目の長所は，そのような同時最適化を行うことができれば，学習中に音響コンテキストが学習データから自動的に決定されていき，私たちが明示的に音響コンテキストを定義する必要がなくなることです．3番目の長所は，分解された隠れ層以外の層は，音響コンテキストに依存せずに共通的に用いることができているため，これらの層のパラメータは全学習データを用いて最適化できる点です．

CADNNによる高速話者適応の実験

CADNNを用いれば，音響モデルを高速にユーザの声に適応することが可能になります．英文新聞ウォール・ストリート・ジャーナルの読み上げ音声の認識に関する単語誤り率（%で表示）を**図4**に示します．単語誤り率は，小さければ小さいほど音声認識の精度が高いことを表しています．ここで用いた私たちのベースライン音声認識システム（図中DNN）の音響モデルは，ReLU（Rectified Linear Unit）活性化関数を用いた5層の隠れ層から成るDNNです．提案するCADNNを用いた音響モデルの構造は，2番目の隠れ層が4つのサブ隠れ層に分解され

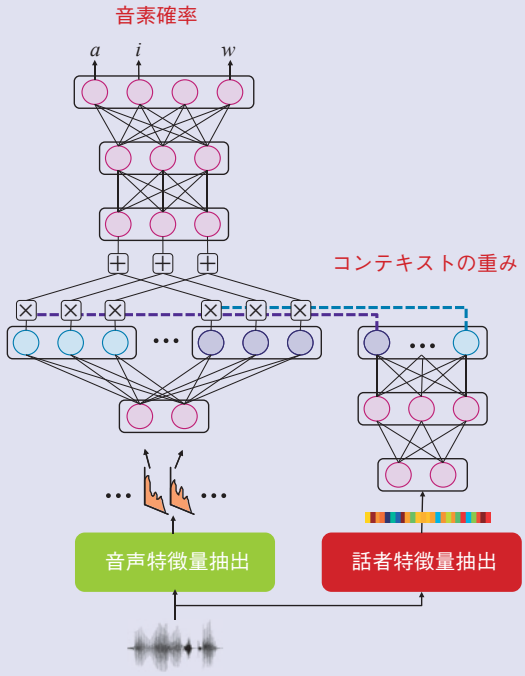


図3　提案CADNNに基づく話者適応

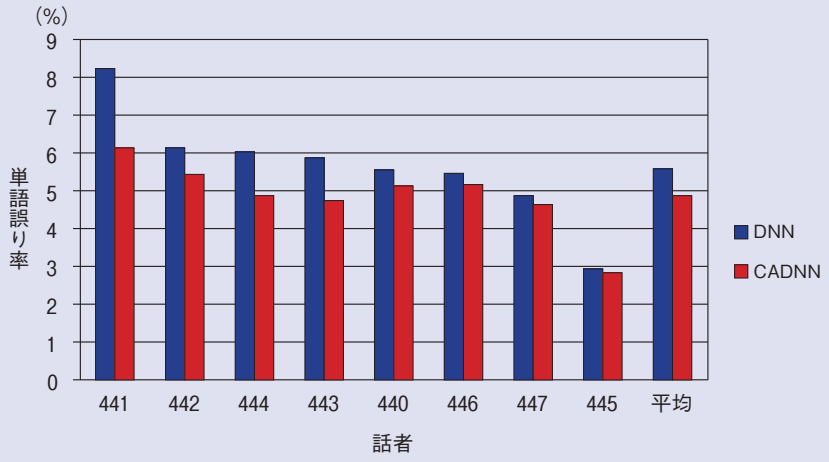


図 4　提案CADNNに基づく話者適応の効果

ている以外は，ベースラインシステムの音響モデルと同じ構成となります．補助的な特徴量としては，話者認識に広く用いられている特徴量を採用し，各発話（10秒未満の音声データ）ごとに計算したものを用いました．

提案するCADNNは，音声認識性能を大幅に改善（ベースラインから約10%）できることが分かります．補助的特徴量の計算には，わずか数秒の音声データがあれば十分で，またその計算のためには書き起こしデータも必要としません．この実験により，CADNNは高速な教師なし話者適応を実現できることが明らかになりました．

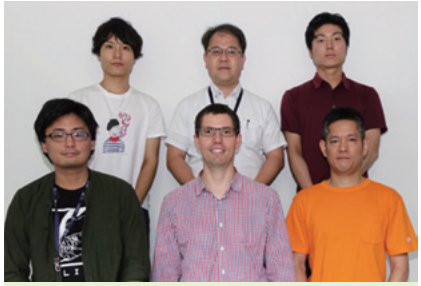
今後の展開

CADNNにより，音響モデルの教師なし高速話者適応の実現可能性が高まりました．より正確な話者表現（話者特徴量）を計算することができれば，将来的には，さらなる性能の向上も見込めるでしょう[4]．また，一発話分のデータから話者特徴量を計算するのではなく，時々刻々と入力される信号から逐次的に話者特徴量を計算していく方法を確立すれば，オンライン適応への道を切り拓くことも可能です[5]．

提案したCADNNの枠組み自体は汎用なものであり，ここで述べた以外の問題の解決にも応用できるものと期待されています．例えば，同様の原理を用いれば，複数の話者の中からターゲット話者の声のみを抽出できることも分かってきています[6]．音声以外の分野においても，DNNの適応が必要な場面は多くあると考えられ，そのような問題にCADNNを応用することも可能となっていくことでしょう．

■参考文献

(1) Y. Kubo, A. Ogawa, T. Hori, and A. Nakamura：“Speech Recognition Based on Unified Model of Acoustic and Language Aspects of Speech,” NTT Technical Review, Vol.11, No.12, 2013.
(2) M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani：“Context adaptive deep neural networks for fast acoustic model adaptation,” Proc. of ICASSP2015, pp.4535-4539, Brisbane, Australia, April 2015.
(3) M. Delcroix, K. Kinoshita, C. Yu, A. Ogawa, T. Yoshioka, and T. Nakatani：“Context adaptive deep neural networks for fast acoustic model adaptation in noisy conditions,” Proc. of ICASSP2016, pp.5270-5274, Shanghai, China, March 2016.
(4) K. Veselý, S. Watanabe, K. Žmolíková, M. Karafiát, L. Burget, and J. H. Černocký：“Sequence summarizing neural network for speaker adaptation,” Proc. of ICASSP2016, pp.5315-5319, Shanghai, China, March 2016.
(5) T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, T. Asami, S. Katagiri, and T. Nakatani：“Cumulative Moving Averaged Bottleneck Speaker Vectors for Online Speaker Adaptation of CNN-based Acoustic Models,” Proc. of ICASSP2017, New Orleans, U.S.A., March 2017.
(6) K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani：“Speaker-aware neural network based beamformer for speaker extraction in speech mixtures,” Proc. of Interspeech 2017, Stockholm, Sweden, August 2017.



（後列左から）木下 慶介/ 中谷 智広/ 樋口 卓哉
（前列左から）苅田 成樹/ Marc Delcroix/ 小川 厚徳

音声認識を，さまざまな人がさまざまな場所で身近な技術として利用できるように，日々研究を進めています．今後は，音声分野への応用を主なターゲットとしつつも，異なる分野への応用も視野に入れながら，検討を進めていきたいと思います．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
メディア情報処理研究部
信号処理研究グループ
TEL　0774-93-5288
FAX　0774-93-5158
E-mail　marc.delcroix@lab.ntt.co.jp