ディープラーニングを用いた実環境における遠隔発話音声処理

近年，音声認識を入力に用いるインタフェースが浸透してきていますが，周囲が騒がしかったり話者とマイクが離れていたりすると，その認識性能は劣化してしまいます．このような場合にも適切に動作する音声認識を実現するためには，さまざまな音声処理要素技術の高精度化が求められます．本稿では，近年注目されているディープラーニングを活用した，音声認識技術およびその前処理である音声強調技術の高度化の取り組みを紹介します．

荒木 章子（あらき しょうこ）/藤本 雅清（ふじもと まさきよ）
吉岡 拓也（よしおか たくや）/Delcroix Marc
Espi Miquel /中谷 智広（なかたに ともひろ）
NTTコミュニケーション科学基礎研究所

遠隔発話音声認識

近年，声で操作できるスマートフォンやタブレット端末が普及し，音声インタフェースの有用性が広く知られるようになってきました．ユーザがマイクの近くで比較的丁寧に話す場合（**図1**(a)）には，その音声は多くの場合正しく認識され，快適な動作が得られます．

一方で，マイクから離れて音声インタフェースを利用したいというニーズも高まっています．例えば，図1(b)のような多人数会話を記録する場合には，テーブル上の端末で会話を認識させたいですし，ロボットやデジタルサイネージなどに話しかける際には，ユーザはマイクからある程度の距離をもって話すでしょう．しかし現状では，マイクから離れて話すと，音声インタフェースの認識性能は大きく低下します．これは，マイクと口とが離れることで周囲の雑音や残響の影響が大きくなることや，ユーザがマイクを意識せず自由な話し方で話すことが要因に挙げられます．NTTコミュニケーション科学基礎研究所では，マイクから離れて話す（遠隔発話）場合にも快適に動作する音声インタフェースを目指して，音声認識と音響処理の研究を進めています．

遠隔発話の場合に音声認識精度が低下する要因を詳しくみると，大きく2つが挙げられます．①空調などの背景雑音や，音声が壁などに反射してからマイクに届く残響の影響が大きくなり，マイクで収録される音声品質が大幅に低下します．また，複数人会話の場合は，ほかの人の声が重なり合って収録されることもあります．②ユーザがマイクを意識せず，自由に話をするようになるため，発音がぼやけたり，言葉を省略したりすることが多くなります．このようなさまざまな要因に対処するためには，雑音や残響・他話者音声の影響を低減する「雑音除去」「残響除去」「音源分離」などの各音声強調技術，そして自由な話し言葉を正確に認識するための「話し言葉音声認識技術」といったさまざまな要素技術が求められます．

遠隔発話音声処理とディープラーニング

私たちは，実環境での遠隔発話音声認識の実現を目指して，前述の各要素技術の研究を行っています．特に近年は，ディープラーニングを適切に活用した各要素技術の高精度化に精力的に取り組んでいます．ディープラーニングとは，深い階層構造を持つニューラ

マイクの近くで丁寧に話す　マイクから離れて自由に話す

(a) 現　状　(b) 目　標

図1　音声インタフェースの現状と目標

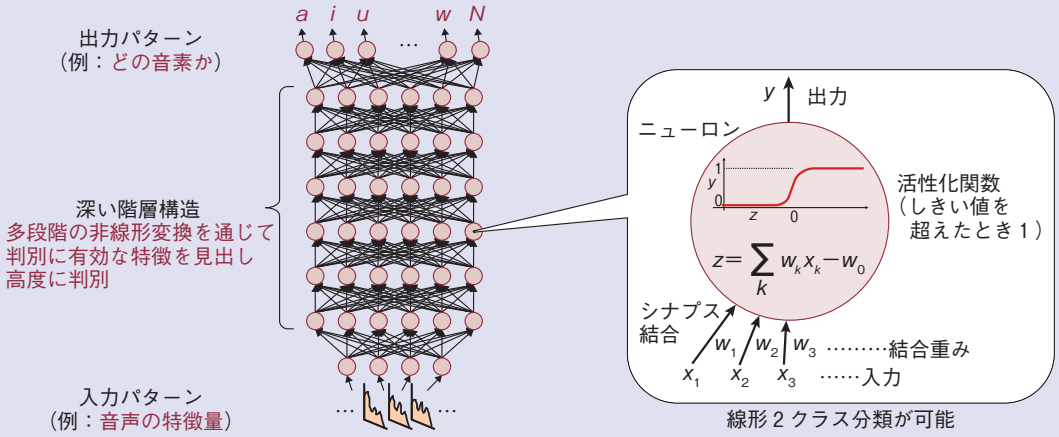


図 2　DNNを用いた音声認識の音響モデル

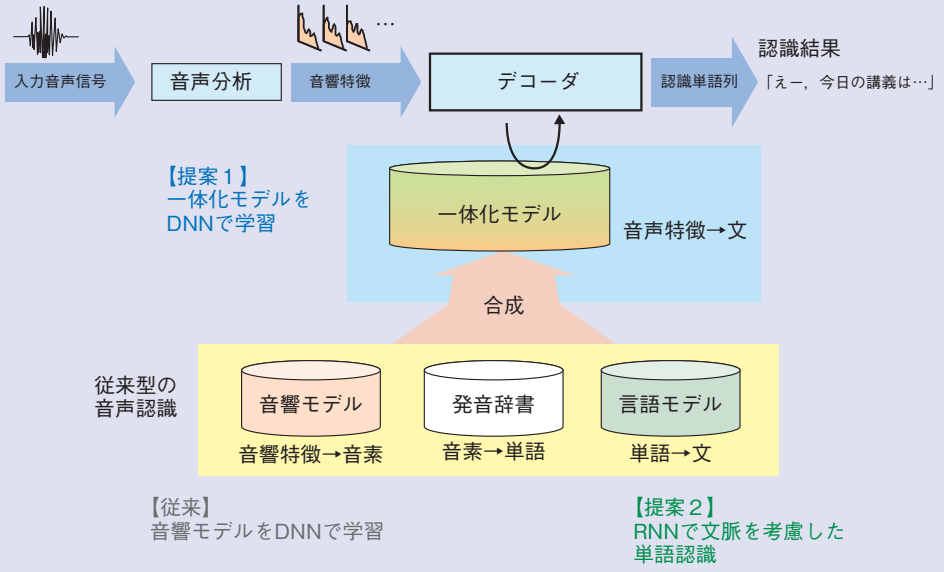


図 3　音声認識の仕組み

ルネットワーク（DNN: Deep Neural Network）を用いた学習の方法で，2011～2012年に画像認識や化合物活性予測をはじめとする幅広い分野で既存技術の性能を次々と上回ったことから，近年注目を集めている技術です．音声認識の分野でも，ディープラーニングを用いた方法がこれまでにない高い性能を達成したことから，さかんに研究されています（**図 2**）．

私たちは2011年から，ディープラーニングを用いた話し言葉音声認識技術の研究を進めており[(1)]，すでに一部をNTTメディアインテリジェンス研究所にてリアルタイム化し商用化しています[(2)]．また，雑音除去などの音声強調技術でも，ディープラーニングをうまく適用することで，その性能を改善できることが分かってきました．次に，私たちが提案しているディープラーニングを用いた音声認識技術と音声強調技術について説明します．

ディープラーニングを用いた音声認識

一般的な音声認識は，**図 3** の黄色背景の部分のように，音響モデル・発音辞書・言語モデルを用い，音声特徴量から音素へ・音素から単語へ・単語から文章へ，それぞれ変換を行います．従来のディープラーニングを用いた音声認識では，音響モデル部分にDNNを適用し，認識性能が格段に向上する

ことが示されました．

音声認識では，上述の３つの各モデルは個別に学習されることが多く，音声と言語の相互作用を反映することが困難です．しかし話し言葉は前述のとおり，声の音としての特徴がなまる，文法的に不正確であるなど，音声認識の難しさは複合的です．そこで私たちは，音声と言語の相互作用を考慮するために，３つのモデルを一体化させ（図３【提案１】），それをDNNで最適化する手法を提案しました[1]．この一体化モデルは話し言葉音声認識に対して高い認識精度を達成することが確認されています．

さらに，言語モデル部分に，DNNの１つであるRNN（Recurrent Neural Network）を用い，これをさらに一体化モデルに組み込むことで，さらなる性能向上が得られます（図３【提案２】）．RNNは，単語の履歴を保持でき，長い文脈を考慮しながらの音声認識が可能になることから，話し言葉音声認識に適します．しかし一般に，文脈の全履歴を保持しながら高速な音声認識をすることは困難です．そこで私たちはここに効率的なアルゴリズムを提案し，リアルタイムで高精度な音声認識を実現しています[3]．

話し言葉の１つである講義音声を認識したときの単語誤り率を**図４**に示します．「DNNなし」はディープラーニング導入以前の認識率です．まずは音響モデルのみにディープラーニングを適用する「DNN音響モデル」の方法で，ディープラーニングの基本的な（そして大きな）効果が確認できます．さらに一体化モデルをDNNで最適化することで（図４の一体化DNN），従来のDNN音響モデルを上回る精度が得られています．また，「RNN言語モデル」の効果も極めて大きく，従来型のDNN音響モデルに比べて４ポイント以上の性能向上が得られました．ディープラーニング手法を適切に用いることで，話し言葉音声認識の大幅な精度向上を達成できます．

ディープラーニングを用いた音声強調

ディープラーニングは音声強調においてもその威力を発揮します．ここでは観測音中の背景雑音を除去する技術として，複数マイクを利用できる場合の方法と，マイクが１個の場合でも高い精度を実現する方法を紹介します．

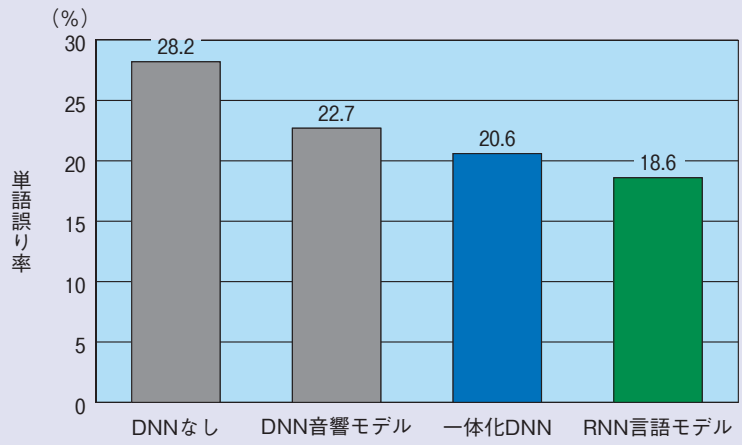


図４　英語講義音声認識の結果

１番目の方法は，DNNで雑音除去音声の特徴量を直接推定（**図５**(a)）します．ここではきれいな音声と，それと発話内容が一致する雑音重畳音声の両方のデータを用いて，雑音重畳音声からきれいな音声への変換をDNNに学習させます．そしてこのDNNに雑音が含まれる音声を入力し，雑音除去音声の特徴量を推定します．従来この方法は，１個のマイクで観測した雑音重畳音声の雑音除去に利用され，複数マイクへの拡張は自明ではありませんでした．これに対し私たちは，複数マイク観測音から推定される特徴量も合わせてDNNに入力することで，よりクリアな雑音除去音声の特徴量を推定できることを示しました．ここで複数マイク観測音からの特徴量としては，各時間周波数スロットにおける音声存在確率（マイクアレイ技術を用いて推定可能）が特に高い性能を与えることが分かってきています[4]．例えば，リビング雑音下での音声認識タスクであるPASCAL CHiME challengeタスクにおける単語誤り率は，単一マイクの場合には10.7%，複数マイク特徴量を併用した場合には8.8%と，その優位性が見て取れます．

加えて私たちは，単一マイクしか利用できない場合でも高い精度を達成する雑音除去法の検討も進めています．私たちのこの方法では，きれいな音声の確率モデルと，音声が含まれない雑音の確率モデルを用意し，このモデルパラメータを用いて雑音除去フィルタを計算します．精度の高い雑音除去フィルタ設計には，きれいな音声モデル・雑音モデルそれぞれを，精度良く推定することがポイントになります．私たちは，きれいな音声モデルの推定にDNNの高い識別性能を活かすことで（図５(b)），より高精度な雑音除去

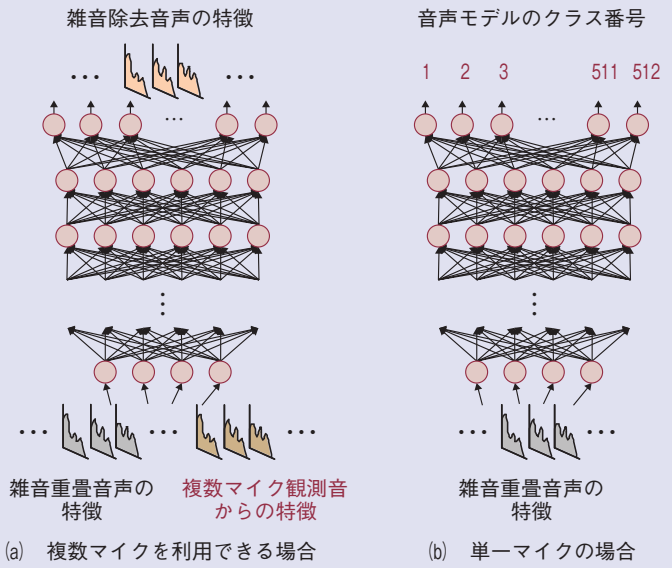


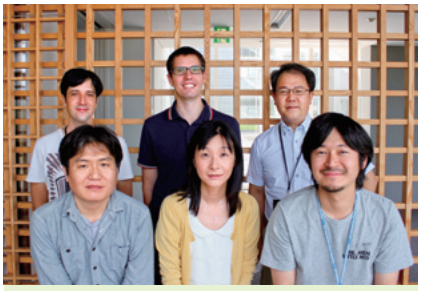
図5　ディープラーニングを用いた雑音除去の例

を達成できることを示しました[5]．すなわち，きれいな音声モデルを多クラスの確率モデルの集合で表し，この中のどのモデルを使えば雑音重畳音声をもっともよくモデル化できるかをDNNで識別させます．この手法により，例えば6種類の雑音下音声データベースAURORA4における単語誤り率は，DNN未使用時に23.0%であるのに対し，DNNを用いた場合は19.6%まで改善します．なおここで，雑音モデルの推定にはDNNを使いません．この理由は，雑音は極めて種類が多いうえ，環境によっては時々刻々変動することから，DNNを学習するための雑音データを十分に確保することが難しいためです．したがって，雑音モデルは学習データが不要な方法を用いてその都度推定し，きれいな音声のモデルはDNNで精度高く選択することで，実際の音環境における多様な雑音に柔軟に対応しながら精度の高い雑音除去を実現することができるのです．

今後の展開

遠隔発話音声処理技術は，音声インタフェースの利用可能性を広げるためのキー技術です．私たちは，遠隔発話音声処理技術の中でも特に，騒がしい環境での多人数会話認識や会話シーン分析が，時代の要請の高い技術であると考えています．この技術は，最近注目を集めている人工知能の音声入力部分，例えばオフィス打合せスペースでの議論の議事録作成やリビングで動作するインテリジェント家電，空港やショッピングセンタでのロボットとの対話などの実現に，大きく寄与するでしょう．そのためには，騒がしい場所で遠隔発話された音声を高精度に認識できることは必須ですが，さらに「誰が話しているのか」を識別したり，音声以外の環境音を認識して「周りで何が起こっているか」を判別したりする技術[6]も重要になります．私たちは，要素技術のさらなる高精度化と，実データを用いた技術評価を並行して行いながら，音声インタフェースの可能性を最大限に拡げていきます．

■参考文献
(1) 久保・小川・堀・中村："音声と言語の一体型学習に基づく音声認識技術，" NTT技術ジャーナル，Vol.25，No.9，pp.22-25，2013.
(2) http://www.ntt-it.co.jp/press/2014/1111/
(3) T. Hori, Y. Kubo, and A. Nakamura："Real-time one-pass decoding with recurrent neural network language model for speech recognition，" Proc. of ICASSP2014, pp.6414-6418, Florence, Italy, May 2014.
(4) S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani："Exploring multi-channel features for denoising-autoencoder-based speech enhancement，" Proc. of ICASSP2015, pp.116-120, Brisbane, Australia, April 2015.
(5) M. Fujimoto and T. Nakatani："Feature enhancement based on generative-discriminative hybrid approach with GMMs and DNNs for noise robust speech recognition，" Proc. of ICASSP2015, pp.5019-5023, Brisbane, Australia, April 2015.
(6) M. Espi, M. Fujimoto, and T. Nakatani: "Detection and classification of acoustic events using multiple resolution spectrogram patch models，" 日本音響学会2014年秋季研究発表会講演論文集，3-8-4，pp.1529-1530，2014.



（後列左から）Espi Miquel/
Delcroix Marc/
中谷 智広
（前列左から）藤本 雅清/ 荒木 章子/
吉岡 拓也

騒がしい場所で多くの人が話すような場面でも，快適に利用できる音声インタフェースの早期実現に向け，より高精度な音声強調と話し言葉音声認識の要素技術の確立に，これからも邁進していきます．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
メディア情報研究部　信号処理研究グループ
TEL　0774-93-5319
FAX　0774-93-5158
E-mail　araki.shoko@lab.ntt.co.jp